Université Montpellier II
Sciences et Techniques du Languedoc

THESE

Pour obtenir le grade de

Docteur de l'Université Montpellier II

Formation Doctorale : Evolution, Ecologie, Ressources Génétiques, Paléontologie
Ecole Doctorale : Systèmes Intégrés en Biologie, Agronomie, Géosciences, Hydro-sciences et Environnement

Présentée et soutenue publiquement le 9 décembre 2008 par

Philippe Cubry

Structuration de la diversité génétique et analyse des patrons de déséquilibre de liaison de l'espèce *Coffea canephora* Pierre ex. Froehner

Devant le Jury composé de :

André Charrier	Professeur émérite, Montpellier Supagro	Président
Dominique Crouzillat	Directeur de recherche, Nestlé	Rapporteur
Sophie Gerber	Chargée de recherche, INRA	Rapporteur
François Anthony	Directeur de recherche, IRD	Examinateur
Domenica Manicacci	Maitre de conférences, Paris XI	Examinateur
Thierry Leroy	Chercheur, CIRAD	Directeur de thèse

Numéro attribué par la bibliothèque :

Université Montpellier II
Sciences et Techniques du Languedoc

THESE

Pour obtenir le grade de

Docteur de l'Université Montpellier II

Formation Doctorale : Evolution, Ecologie, Ressources Génétiques, Paléontologie
Ecole Doctorale : Systèmes Intégrés en Biologie, Agronomie, Géosciences, Hydro-sciences et Environnement

Présentée et soutenue publiquement le 9 décembre 2008 par

Philippe Cubry

Structuration de la diversité génétique et analyse des patrons de déséquilibre de liaison de l'espèce *Coffea canephora* Pierre ex. Froehner

Devant le Jury composé de :

André Charrier	Professeur émérite, Montpellier Supagro	Président
Dominique Crouzillat	Directeur de recherche, Nestlé	Rapporteur
Sophie Gerber	Chargée de recherche, INRA	Rapporteur
François Anthony	Directeur de recherche, IRD	Examinateur
Domenica Manicacci	Maitre de conférences, Paris XI	Examinateur
Thierry Leroy	Chercheur, CIRAD	Directeur de thèse

Remerciements :

Voici sans doute la partie la plus libre de ce manuscrit, mais pas nécessairement la plus évidente à écrire. Je vais donc essayer de n'oublier personne dans la longue liste que je dois dresser.

Mes premiers remerciements iront bien entendu à ma famille, en particulier mes parents et mon frère, ainsi qu'à Géraldine, qui partage ma vie et me supporte depuis quelques temps déjà. Ces personnes m'ont toujours soutenu y compris dans mes moments d'égarement (et il y en a eu de nombreux), un énorme merci donc, et un grand « je vous aime ».

Passons à l'historique de mon arrivée au CIRAD. Tout à commencé un beau jour de maîtrise où j'ai dû rechercher un stage, ce fut sur le bananier, m'initiant aux problématiques spécifiques abordées par le CIRAD et me faisant découvrir un laboratoire dynamique et intéressant, merci à Claire Billot et Isabelle Hyppolite qui m'ont offert ce sujet et m'ont fait découvrir ce labo.

Vint ensuite le DEA, et rebelote, un nouveau stage. C'est à ce moment là que j'ai découvert le café et son univers si particulier. Merci à Magali Dufour, qui m'a proposé un sujet de stage particulièrement intéressant qui a servi de prémices au travail présenté ici. Merci également à David Pot pour ses discussions d'alors et sa bataille avec les gels de carto, et puis pour tout ce qui a suivi également. C'est également au cours de ce stage que j'ai pu apprendre à connaître et à travailler avec Fabien De Bellis, merci beaucoup Fab pour tous les conseils prodigués, les discussions quasi philosophiques sur les recherches en cours et à venir et puis pour tout le reste également, même si je n'ai toujours pas réussi à te trainer à Mauguio pour voir à quoi peut ressembler une discussion d'article, je ne désespère pas d'y arriver un jour, je pense que ça te plairait ! Toujours au cours du DEA, j'ai connu l'alors chef de l'équipe café, devenu depuis mon directeur de thèse. Même si nous ne sommes pas toujours d'accord et que nos caractères bien trempés se sont quelquefois affrontés, mais je vous rassure jamais méchamment, je pense sincèrement avoir beaucoup appris et je remercie donc à ce titre monsieur Thierry Leroy. Puisque nous en sommes toujours à cette période d'avant-thèse, une camarade de galère était déjà présente, Sophie, merci pour tout et courage pour ta soutenance dans, euh, quelques jours après moi.

Un merci particulier aux Julie (x3) qui se reconnaîtront pour l'ensemble de toutes les discussions, scientifiques, philosophiques et autres que nous avons pu échanger. Merci aux thésards du bâtiment que je n'ai pas déjà cité, Aurélie, Vincent, Benoît, unis dans la même galère, et bon courage au suivants, au compte desquels Daniel, mon successeur à la torture des

Licors. Parlant des Licors, tout naturellement un grand merci à Ronan Rivallan, responsable de la plateforme de génotypage, qui réalise un travail titanesque mais toujours avec bonne humeur, bon courage pour la suite Roro !

Bien sûr une pensée également pour Jean-Louis Noyer, chef de l'équipe SRG, qui m'a permis de me mettre le pied à l'étrier dans l'enseignement, un grand merci à toi pour cette chance.

Mes remerciements iront également à mes colocataires de bureau au cours du temps passé au CIRAD, j'en oublierais surement mais je citerai Mathilde, Michel, Roger, Giang, Carole, Pratap parmi ceux-ci.

Elargissons un peu notre horizon à l'ensemble du bâtiment 3 du Cirad, merci donc à tout le personnel permanent et temporaire qui est passé par là au cours des 4 ans et demi qui viennent de passer.

Après les personnes du labo, ceux de l'université ou des autres instituts de recherche, merci à Isabelle Olivieri et Jacques David qui m'ont transmis leur enthousiasme de la recherche. Merci également aux membres de mon comité de thèse, Loïc Le Cunff, Marie-Hélène Muller, Mathilde Causse, Patrice This pour leurs conseils et leur attention. Merci aux deux personnes qui ont accepté de passer du temps à juger ce travail, Sophie Gerber et Dominique Crouzillat, ainsi qu'aux autres membres du jury, Domenica Manicacci, André Charrier et François Anthony, pour leurs conseils précieux. Un petit mot particulier pour François, qui m'a permis de mieux comprendre le caféier et qui m'a transmis de nombreuses données sur les nanas (non, pas celles que vous croyez...) et qui a été d'une grande aide durant la rédaction de mon premier article.

Passons maintenant aux personnes qui me sont proches en dehors du travail. Un grand merci à Betty, Denis et Emmanuel pour me supporter depuis maintenant quelques (longues) années et pour tout ce qu'ils m'ont appris, mon affection leur est acquise. Merci également aux membres des chœurs A et B d'Oratorio, en particulier Marina et A2. Thierry et Aurélie, merci pour votre amitié et encore félicitations pour la magnifique petite fille que vous venez d'avoir, un bisou à Coline ! A Danielle, Jacques, Ghislaine, Elisabeth et tous les autres avec qui j'apprends énormément en tant que responsable et bénévole d'association. Une pensée également à Edouard et Ella, qui nous suivent depuis le début.

Merci aux différents membres du CFM, entité totalement auto constituée, mais qui depuis maintenant 12 ans continue à exister dans notre quotidien, on se verra le 31 les potes et ce sera l'orgie. Plus sérieusement un grand merci à Olivier et Sandrine, auquel j'ajouterais B-Scott, pour les moments passés dans le Var et tout le reste, à David, toujours prêt à toutes les

plus grandes conneries possibles, j'espère que tu vas réussir à les faire tes économies !! Merci à Cécile et Pascal, et maintenant à Clara que je n'ai toujours pas eu le plaisir de voir, à Laure et Pierre, à Nico (non tu n'es pas une chochotte, et oui les dents c'est très important), Jules et Agnès, à Guitou et sa prof de Julie. Un merci tout particulier à Virginie et Olivier, particulièrement pour m'avoir poussé à venir vous voir du côté de Nice à une période pas très agréable pour moi et puis pour tous les autres moments (z'inquiétez pas, cette année on se la fait la vieille).

Ainsi s'achèvera cette interminable liste, et malgré sa longueur, j'en ai très certainement encore oublié, alors que ceux que je n'ai pas cités veuillent bien m'en excuser.

SOMMAIRE

Chapitre 1 : Introduction et synthèse bibliographique

La sélection pour l'amélioration de la qualité et de la productivité des plantes a débuté dès la préhistoire avec la domestication. Cette dernière a donné lieu à de nombreux goulots d'étranglement successifs caractérisés par une baisse drastique de la diversité dans les compartiments cultivés, d'abord avec les variétés traditionnelles, puis encore plus prononcée avec les variétés élites. Le développement de la génétique quantitative, de la génétique des populations et plus récemment de la génétique moléculaire a permis de donner un cadre théorique à la sélection que l'on faisait de manière empirique jusque là. Ces avancées technologiques et conceptuelles ont permis la mise en place de schémas de sélection raisonnés et contrôlés. La génétique d'association permet aujourd'hui d'envisager des manières d'optimiser ces schémas en y intégrant par exemple une part de Sélection Assistée par Marqueurs (SAM). Nous allons dans ce premier chapitre résumer l'état des connaissances sur la diversité des caféiers cultivés, en s'intéressant particulièrement à l'espèce *Coffea canephora*, avant de donner quelques informations sur le schéma de sélection actuellement en cours en République de Côte d'Ivoire (RCI). Nous exposerons ensuite le principe de la recherche d'association marqueur/caractère d'intérêt et l'aspect théorique sur lequel il repose, le déséquilibre de liaison.

Le genre *Coffea* et les caféiers cultivés

Le genre *Coffea* comprend une centaine d'espèces, il est endémique des régions tropicales et sub-tropicales d'Afrique et de Madagascar (Davis & Stoffelen, 2006). Parmi cette importante diversité, seulement deux espèces sont cultivées de manière significative, *Coffea arabica* L. et *Coffea canephora* Pierre ex Froehner. *C. arabica* et *C. canephora* représentent l'immense majorité de la production mondiale et sont les seules espèces incorporées dans les statistiques de l'ICO (international coffee organisation, www.ico.org). Selon le bulletin de septembre 2008 de cette organisation, *C. arabica* représente environ 61,5% de la production mondiale, *C. canephora* représentant les 38,5% restants. Si *C. canephora* est une espèce diploïde, allogame stricte, à 2n=2X=22, *C. arabica* est quant à elle une allotétraploïde autocompatible à 2n=4X=44.

Si l'on considère l'histoire de la culture des caféiers, celle-ci est relativement complexe et diffère selon les espèces. *C. arabica* est reconnue comme cultivée depuis plusieurs siècles et ses migrations successives sont bien documentées. L'histoire de la culture de *C. canephora* ne commence en revanche qu'il y a tout au plus 150 à 200 ans.

C. arabica

L'histoire de la culture de *C. arabica* remonte à plusieurs siècles. L'introduction de cette espèce au Yémen à partir de son centre de diversité originel (Sud ouest de l'Ethiopie) a pu se dérouler il y a 500 à 1500 ans. Les vagues successives de diffusion de *C. arabica* se sont ensuite déroulées à partir du Yémen, à compter de la découverte en Europe de la dégustation de la boisson produite à partir du XVII^{e} siècle. Après ces vagues de diffusion, réalisées à partir de quelques graines, la diversité génétique de base des variétés cultivées est extrêmement étroite, avec seulement quelques génotypes (Figure 1.1). La variété Typica est, par exemple, issue d'un seul plant du jardin botanique d'Amsterdam (Anthony *et al.*, 2002).

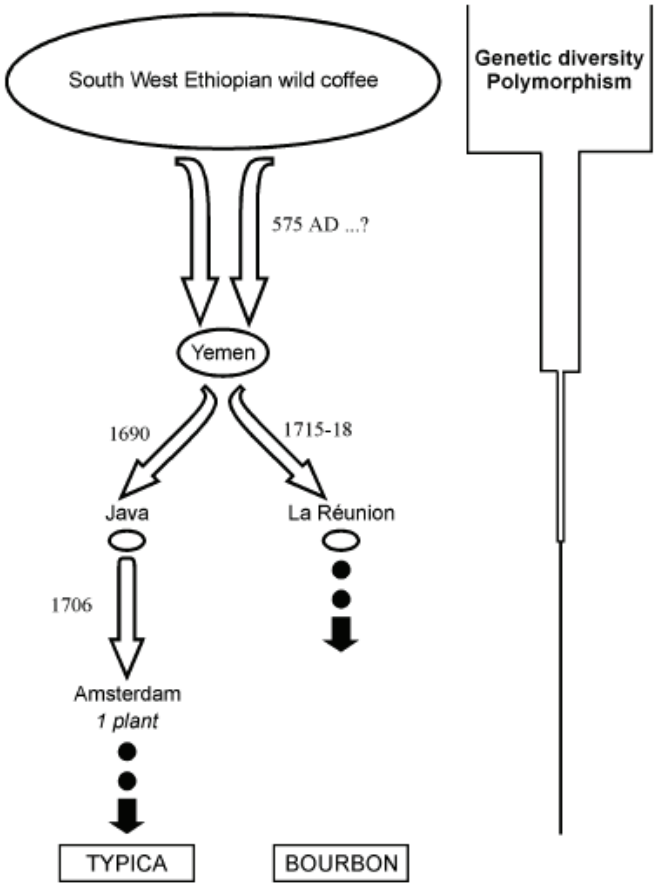


Figure 1.1 : Historique de la culture et des migrations successives de *C. arabica*, en relation avec la diversité génétique du compartiment cultivé. Tiré de Anthony *et al.* (2002)

Cette base génétique étroite rend les cultures fortement sensibles aux atteintes parasitaires. Un certain nombre de prospections réalisées dans les centres de diversité primaire (Ethiopie) et secondaire (Yémen) ont permis d'enrichir les collections et de fournir une variabilité pour la sélection. Néanmoins la diversité de l'espèce reste globalement faible et le recours à des espèces voisines, notamment *C. canephora*, est utile pour la sélection (Charrier & Eskes, 2001; Eskes & Leroy, 2004) (Figure 1.2). Les principales cibles de la sélection pour *C. arabica* sont les résistances aux maladies (*Hemileia vastatrix*, anthracnose des baies), aux nématodes (*Meloidogyne* spp. et *Pratylenchus* spp.) et aux insectes (scolyte des baies, mineuse des feuilles).

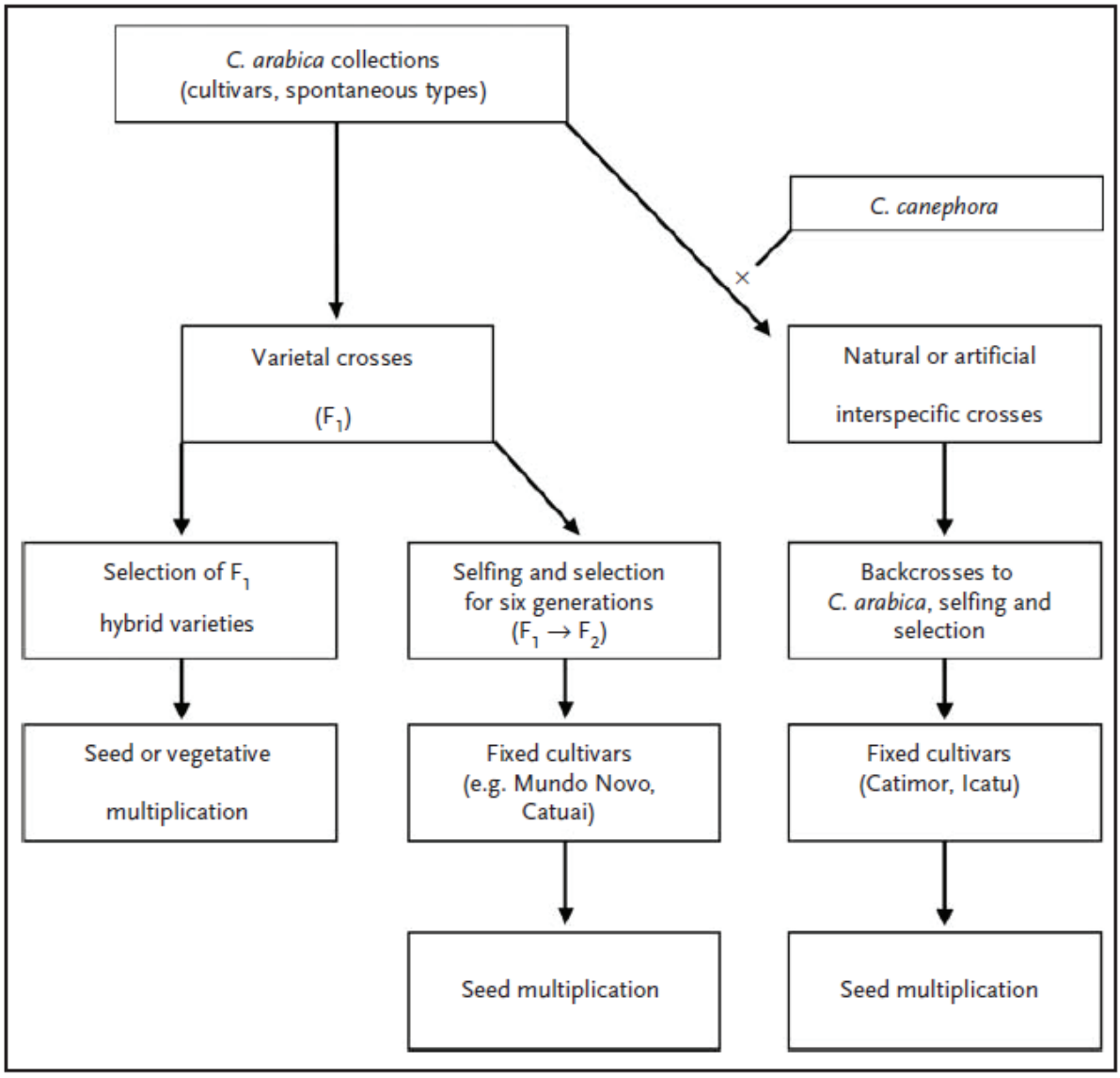


Figure 1.2 : Schéma de sélection applicable à *C. arabica*. Tiré de Eskes & Leroy (2004)

C. canephora

La mise en culture ou l'utilisation de *C. canephora*, bien qu'ayant pu avoir commencé au XVIII^{e} siècle pour des aspects rituels, notamment en Ouganda, a surtout été développée au début du XX^{e} siècle. A cette époque, d'importantes attaques parasitaires des cultures de *C. arabica* ont conduit à tester des espèces de remplacement plus résistantes. Après des essais plus ou moins concluants avec l'espèce *C. liberica*, *C. canephora* a été retenue pour sa vigueur et sa tolérance aux maladies. Des travaux de sélection ont débuté à cette époque à Java en Indonésie à partir d'introduction dès 1900 du Congo belge puis vers 1915 du Congo-Brazzaville ou du Gabon (Cramer, 1957).

La diversité génétique de *C. canephora* a été étudiée successivement à l'aide de marqueurs isoenzymatiques (Berthaud, 1986; Montagnon *et al.*, 1992) puis par RFLP (Dussert *et al.*, 1999). Ces études ont montré à la fois la diversité très importante de cette espèce et sa forte structuration. Berthaud a ainsi décrit deux grands groupes de *C. canephora*, l'un correspondant aux génotypes de Guinée et Côte d'Ivoire, les Guinéens, et l'autre regroupant les génotypes d'Afrique centrale, les Congolais. Montagnon et Dussert ont ensuite précisé cette structure en subdivisant successivement les Congolais en 2 (SG1 et SG2) puis 4 groupes (SG1, SG2, B et C). Cette diversité génétique se superpose à une diversité phénotypique importante (Montagnon *et al.*, 1992; Montagnon & Leroy, 1993; Montagnon, 2000) permettant d'envisager des potentialités de gains génétiques importants.

Contrairement à *C. arabica*, la diversité des génotypes cultivés de *C. canephora* est importante, en raison de son caractère allogame et de la multitude des centres d'origine et de diversification. De plus la courte durée de l'histoire de la culture pour cette espèce a limité le brassage génétique et le goulot d'étranglement de la domestication, résultant en une absence à l'heure actuelle de syndrome de domestication. En République de Côte d'Ivoire, l'utilisation de génotypes locaux et introduits qui se sont intercroisés a aussi contribué à créer et maintenir une forte diversité dans le compartiment cultivé.

Les différents programmes de recherche successifs sur *C. canephora* ont été décrits dans plusieurs articles (Montagnon *et al.*, 1998a; Montagnon *et al.*, 1998b). La sélection a d'abord été réalisée à l'intérieur de chaque groupe de diversité, puis par l'utilisation d'hybrides intergroupes. Aujourd'hui le schéma de Sélection Récurrente et Réciproque (SRR) existant en Côte d'Ivoire (Figure 1.3) repose à la fois sur la sélection intragroupe et l'hybridation intergroupe (Leroy *et al.*, 1993; Leroy *et al.*, 1994; Leroy *et al.*, 1997; Dussert *et*

al., 1999). Ce schéma pourrait être amélioré par l'apport des études d'associations et la sélection assistée par marqueurs (Leroy, communication personnelle).

Les principales cibles de la sélection sur *C. canephora* sont la recherche de résistances, l'amélioration de la qualité du produit et la tolérance à la sécheresse.

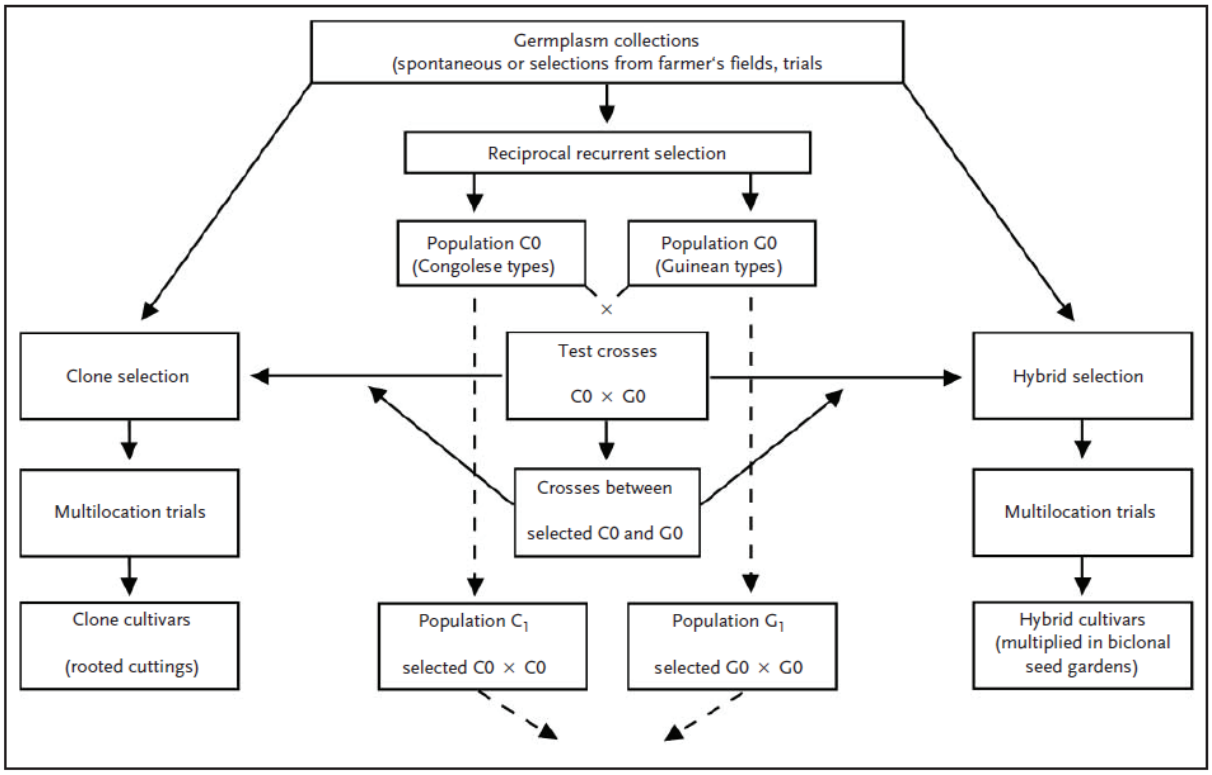


Figure 1.3 : Schéma de sélection applicable à *C. canephora*. Tiré de Eskes & Leroy (2004)

C. canephora est une espèce relativement facile à travailler en comparaison de *C. arabica*, en particulier à cause de son caractère diploïde et de sa très forte diversité génétique et phénotypique. Les connaissances accumulées sur *C. canephora* et sa place phylogénétique permettent d'espérer étendre les résultats obtenus sur cette espèce aux autres espèces du genre, et notamment *C. arabica*, la plaçant en situation d'espèce ressource pour le genre *Coffea*.

La recherche d'association marqueur/caractère

Si la génétique quantitative est apparue au début du XIX° siècle, sous l'influence de Ronald A. Fisher, son objectif, étudier les déterminismes génétiques des caractères complexes, a connu un fort bouleversement avec l'avènement récent des marqueurs moléculaires et l'amélioration continuelle des techniques de génotypage. Nous en sommes

aujourd'hui à un point où le coût de séquençage ou de marquage moléculaire devient un élément tout à fait gérable et non limitant pour la plupart des études menées sur les plantes.

Les marqueurs moléculaires semblent être les outils ultimes pour la recherche d'associations entre les variations des caractères d'intérêt agronomique observés chez les plantes et les polymorphismes génétiques à l'origine de ces variations. Ce type d'étude a été fortement développé au cours des dernières décennies, d'abord par des approches que nous qualifierons de cartographie génétique « classique » (recouvrant le terme anglais de « linkage analysis ») associées à des recherches de QTL (locus impliqué dans la variation d'un caractère quantitatif), puis par des approches de génétique d'association.

Ces 2 types d'approches reposent sur la même hypothèse d'héritage partagé des polymorphismes fonctionnels et des polymorphismes ADN avoisinants (notion de déséquilibre de liaison développée ci-dessous).

Qu'est-ce que le DL, et comment le mesurer?

Le déséquilibre de liaison, ou déséquilibre de phase gamétique, se définit littéralement comme l'écart à l'association aléatoire entre allèles à des locus différents. En d'autres termes, les fréquences gamétiques des 2 locus considérés ne seront pas égales aux produits des fréquences alléliques correspondantes à ces 2 locus, s'écartant d'une hypothèse aléatoire de recombinaison des allèles au hasard. Si cette notion est assez ancienne (Veyrieras, 2006) la formalisation mathématique de celle-ci date d'environ une cinquantaine d'année, avec le développement des premiers marqueurs isoenzymatiques (Lewontin & Kojima, 1960). Cette formalisation est présentée en Encadré 1.1.

Encadré 1.1 : Formalisation du Déséquilibre de liaison (DL) ou Déséquilibre de Phase Gamétique : mesure *D* de Lewontin et Kojima.

Considérons 2 locus à 2 allèles : A/a et B/b, sous l'hypothèse d'une recombinaison aléatoire des allèles entre des locus différents on obtient le tableau suivant :

	p(B)	p(b)
p(A)	p(AB) = p(A)p(B)	p(Ab) = p(A)p(b)
p(a)	p(aB) = p(a)p(B)	p(ab) = p(a)p(b)

Dans le cas où apparaît un Déséquilibre de liaison (DL) les égalités présentées ci-dessus ne sont plus valables, on introduit alors une statistique, *D*, pour mesurer l'écart entre les fréquences gamétiques et les produits des fréquences alléliques correspondantes :

	p(B)	p(b)
p(A)	p(A)p(B) + D	p(A)p(b) - D
p(a)	p(a)p(B) - D	p(a)p(b) + D

Par conséquent l'équation proposée par Lewontin est : $D = p_{AB} - p_A p_B$

Cette mesure varie au maximum entre -0,25 et 0,25 et est dépendante des fréquences alléliques aux locus considérés.

La sensibilité très importante de cette première mesure aux fréquences alléliques ne permettant pas la comparaison entre des locus différents, des normalisations de *D* ont donc été proposées. Parmi différentes mesures, les plus usitées en génétiques des populations et dans la majorité des travaux actuels sont *D'* (Lewontin, 1964) et r² (Hill & Robertson, 1968). Ces mesures sont présentées en Encadré 1.2.

Encadré 1.2 : Normalisations de *D* permettant des comparaisons entre des locus différents.

- *D'* (Lewontin, 1964)

$$D' = \frac{|D_{AB}|}{\max(D_{AB})}$$

Dans cette équation max(D_{AB}) est défini comme :

$$\max(D_{AB}) = \min[p_A p_B, p_a p_b] \text{ si } D_{AB} < 0$$

$$\max(D_{AB}) = \min[p_A p_b, p_a p_B] \text{ si } D_{AB} > 0$$

Cette mesure du DL varie entre 0 et 1 quelques soient les fréquences au locus considéré.

- r² (Hill & Robertson, 1968)

$$r^2 = \frac{(D_{AB})^2}{p_A p_a p_B p_b}$$

Cette mesure varie entre 0 et 1 et est assimilable au carré du coefficient de corrélation entre les états alléliques.

Ces 2 mesures, bien que reliées et pouvant être écrite l'une en fonction de l'autre sous la forme : $r^2 = (D')^2 \times \frac{f_{(a)} f_{(B)}}{f_{(A)} f_{(b)}}$ si $D \geq O$ et $f_{(A)} \geq f_{(B)}$ (Wang *et al.*, 2005), ont des comportements et des interprétations biologiques différents. En effet *D'* ne va mesurer que les évènements de type recombinaison entre 2 locus alors que r² va résumer à la fois les recombinaisons et les mutations. Si *D'* est plus précis que r² pour la mesure des recombinaisons, r² permettra d'avoir une appréciation de la manière dont sont corrélés les allèles avec des marqueurs ou éventuellement des caractères phénotypiques (Flint-Garcia *et al.*, 2003). Un exemple de la différence de comportement de *D'* et r² est donné en Figure 1.4.

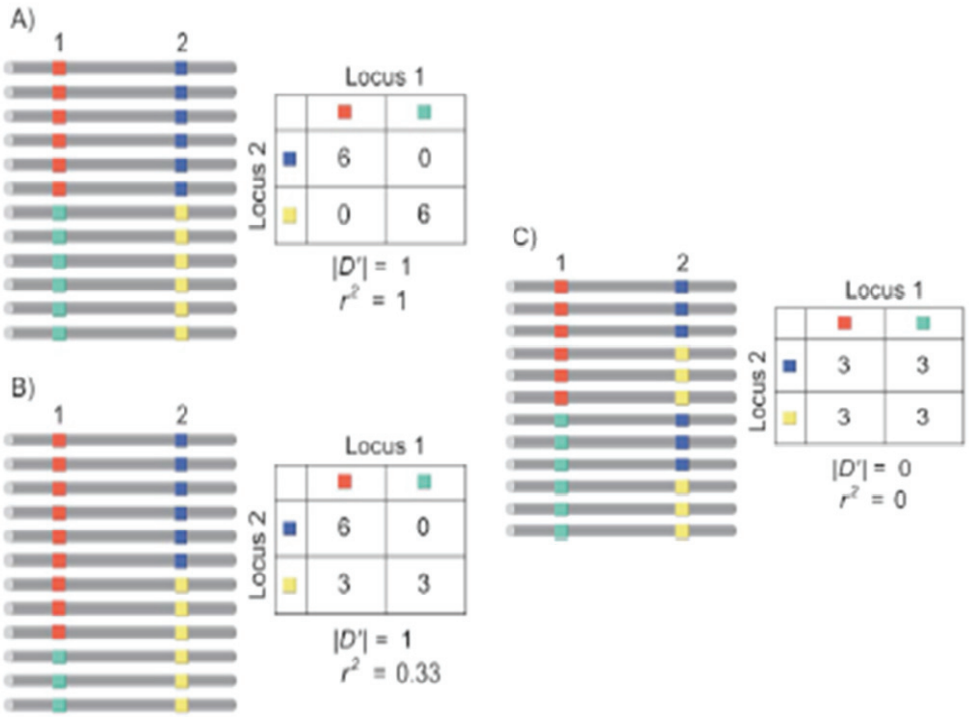


Figure 1.4 : Valeurs de *D'* et r² pour 3 cas différents de déséquilibre entre 2 marqueurs, en (A) le DL est dit total, la connaissance de l'allèle à un locus permet de prédire l'allèle au second locus, en (B) le DL est dit complet, on n'observe que 3 types de gamètes sur les 4 possibles, en (C) équilibre d'association. Les valeurs de *D'* et r² sont différentes dans le cas de DL intermédiaire ou complet. D'après Flint-Garcia *et al.* (2003)

Ces différences de comportement entrainent une différence dans l'utilisation de ces statistiques. Si *D'* est couramment utilisé pour étudier l'histoire des populations et des recombinaisons, r² sera préféré pour les études d'associations car il reflète la manière dont pourront être corrélés les allèles et les caractères d'intérêt. Il faut cependant noter que ces deux mesures sont sensibles aux faibles effectifs et aux allèles rares. Cette sensibilité semble néanmoins plus importante pour *D'* (Flint-Garcia *et al.*, 2003; Wang *et al.*, 2005; Zhu *et al.*, 2008).

Il est important de noter que dans tous les cas le DL mesuré est une association statistique entre 2 locus. Ce déséquilibre pourra être modelé et influencé par l'ensemble des facteurs évolutifs agissant sur les populations.

Les facteurs modelant le déséquilibre de liaison des populations et des espèces

La mutation comme force de création du DL

La mutation est le phénomène évolutif qui va créer le polymorphisme qui pourra être en déséquilibre, c'est donc le moteur de la création du DL. Le DL aux alentours des nouveaux polymorphismes sera important (voir Figure 1.5) tant qu'il ne sera pas dissipé par la recombinaison. Néanmoins un fort taux de mutation va diminuer le DL global.

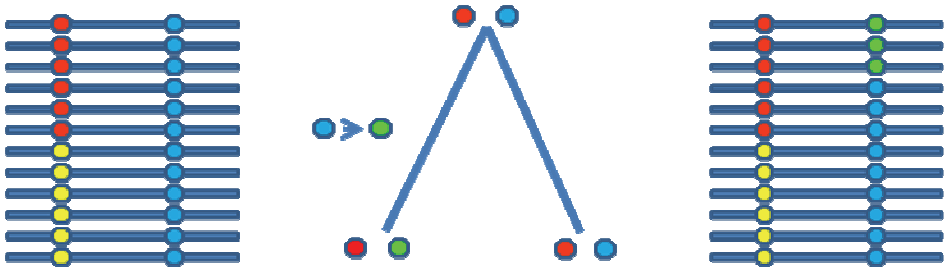


Figure 1.5 : Création de déséquilibre par la mutation, à gauche la population initiale, à droite la population après l'évènement de mutation présenté dans l'arbre. Adapté de Flint-Garcia *et al.* (2003)

La sélection

La sélection sur un locus va pouvoir entrainer une augmentation locale du déséquilibre de liaison par autostop génétique. Ce phénomène s'accompagne généralement d'une baisse simultanée de la diversité aux alentours du locus sélectionné (Figure 1.6).

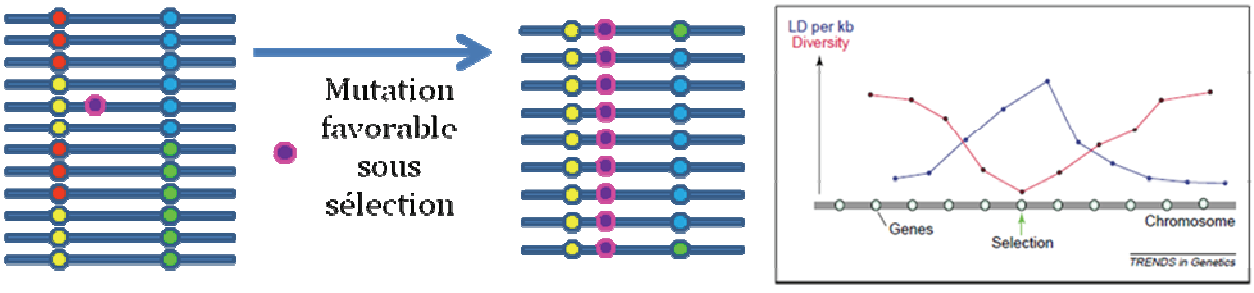


Figure 1.6 : Effet de la sélection. S'il apparaît dans la population initiale (à gauche) une mutation favorable, celle-ci va peu à peu envahir la population, entrainant avec elle les polymorphismes proches. Adapté de Rafalski & Morgante (2004)

Recombinaison ou comment « casser » le DL

La recombinaison est le facteur majeur de dissipation du déséquilibre de liaison. En l'absence de toute autre force évolutive, la décroissance au cours du temps du DL est une fonction du taux de recombinaison. Au plus 2 marqueurs seront proches, et donc au moins le taux de recombinaison est élevé, au moins le DL diminuera à la génération suivante. Ceci est décrit par la relation suivante : $D_{t+1} = D_t(1-\theta)$ soit au bout de t générations : $D_t = D_0(1-\theta)^t$, avec θ le taux de recombinaison entre les 2 locus considérés (Figure 1.7).

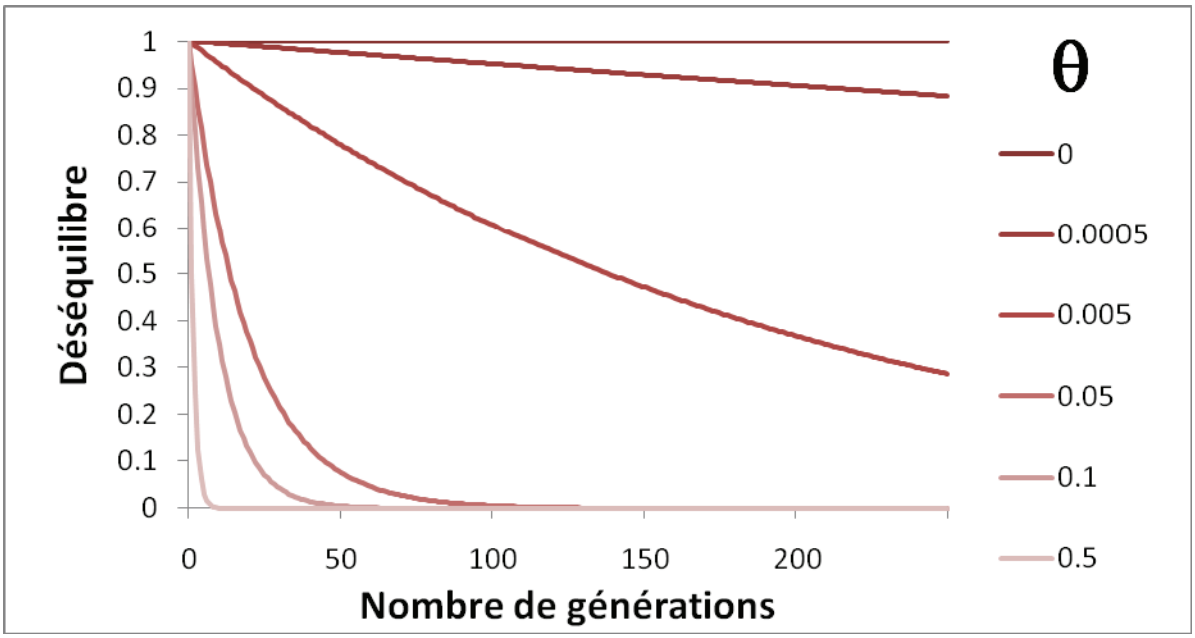


Figure 1.7 : Décroissance du DL en fonction du nombre de générations et du taux de recombinaison θ

On peut montrer que dans une population de taille finie, en l'absence de toute autre force évolutive, l'espérance de r² à l'équilibre entre mutation et dérive est une fonction directe de la taille efficace et du taux de recombinaison :

$$E(r^2) = \frac{1}{1+4N_e c}$$

Avec N_e l'effectif efficace de la population et $c = \theta_{loc} d$ avec θ_{loc} le taux de recombinaison local et d la distance entre les 2 locus considérés.

Par conséquent on peut dériver de ces relations que si la mise en place du DL n'est pas forcément une fonction de la distance physique, son maintien est, lui, une fonction directe de celle-ci à travers le taux de recombinaison. Cela signifie également que le déséquilibre va être

variable le long du génome, le taux de recombinaison étant variable selon les régions. De cette variation on peut déduire un modèle en « île » du DL (Rafalski & Morgante, 2004). Des blocs d'haplotypes conservés peuvent donc être séparés par des points chauds de recombinaison. La variation du taux de recombinaison peut être très schématiquement superposée au contenu en gène de la région considérée. En effet des études ont montré que les zones de fort contenu en gènes sont généralement des zones de fortes recombinaisons (voir Figure 1.8). On s'attend donc à trouver des blocs de DL plutôt dans des régions contenant peu de gènes.

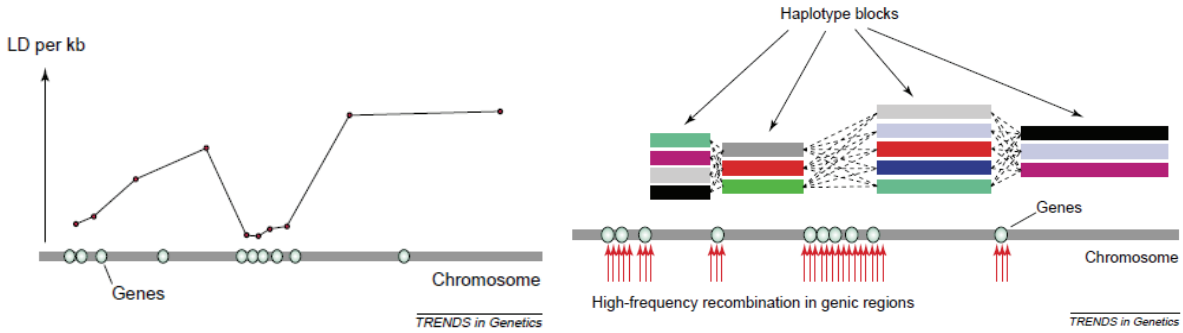


Figure 1.8 : Variation de l'intensité du DL en fonction du contenu en gène et modèle en île, blocs d'haplotypes séparés par des zones de fortes recombinaisons. Tiré de Rafalski & Morgante (2004)

Les facteurs démographiques, la dérive génétique, les goulots d'étranglement

De la relation précédente (espérance de r^2), il découle que les modifications d'effectifs efficaces des populations, donc les facteurs démographiques et le système de reproduction, vont influencer directement sur l'étendue du DL. Les populations de faibles effectifs vont, par dérive génétique, avoir tendance à la perte continue d'allèles rares, ce phénomène ayant tendance à augmenter le DL. De la même manière, un goulot d'étranglement va engendrer de manière temporaire un DL très élevé qui subsistera tant que la recombinaison ne le dissipera pas. Enfin le système de reproduction va avoir une influence primordiale sur les patrons de DL en agissant indirectement sur la taille de l'effectif efficace et l'efficacité de la recombinaison. On s'attend donc à trouver un DL plus important chez les espèces à reproduction autogame que chez les espèces à reproduction allogame (Flint-Garcia *et al.*, 2003). De la même manière on trouvera vraisemblablement un DL plus important dans les lignées élites que dans les cultivars traditionnels et le compartiment sauvage (Rafalski & Morgante, 2004).

Problème de la structure et de la détection de « faux » DL (DL non physique)

Lorsqu'on désire étudier des corrélations entre marqueurs moléculaires et caractères quantitatifs, nous souhaitons nous intéresser à un DL seulement entre marqueurs physiquements liés. Nous pouvons donc identifier deux types de DL, un DL local, entre marqueurs physiquement liés, et un DL au niveau du génome entier. La dérive (par augmentation globale du DL) et le mélange de populations (voir Figure 1.9) vont augmenter ce DL que nous pourrons qualifier de « génomique ».

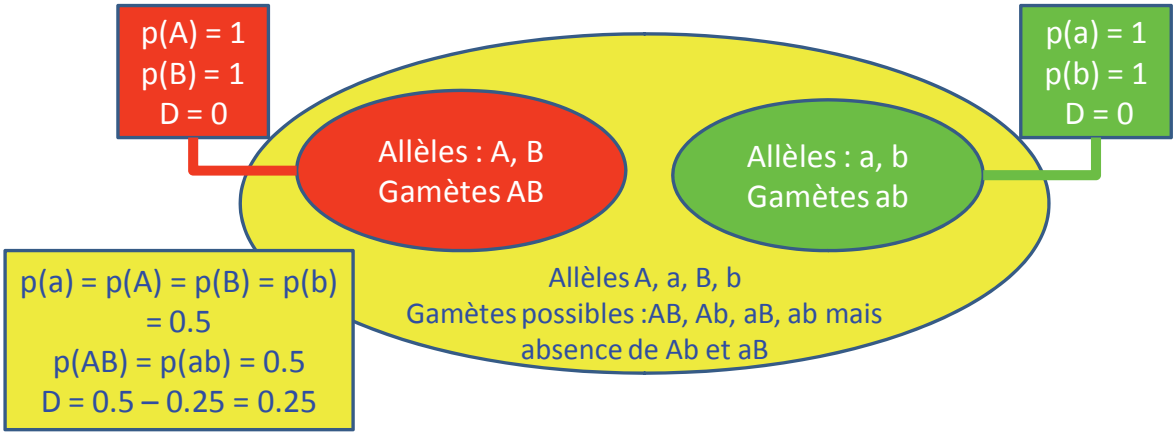


Figure 1.9 : Effet du mélange de population sur le DL. 2 populations (rouge et verte) ne présentant pas de déséquilibre sont mélangées dans l'échantillon jaune et génèrent par ce simple mélange du déséquilibre entre les 2 locus considérés.

Ce problème de la structure génétique sur la création de DL génomique est la principale source d'erreur dans les études d'associations. Nous verrons par la suite que plusieurs modèles statistiques ont été développés pour tenir compte de ce paramètre.

Les principales différences entre cartographie « classique » et études d'association

Les principales différences entre cartographie « classique » et études d'associations reposent sur le type de populations et le nombre de recombinaisons possibles. Dans les approches de cartographies, seulement quelques générations de méiose peuvent être mises en jeu (et par conséquent peu de recombinaisons) et la descendance est parfaitement connue, les populations de cartographie étant issus de croisements contrôlés. De plus la diversité à laquelle on s'adresse est relativement limitée et représente au plus 4 allèles pour une espèce

diploïde en utilisant des populations issues d'un croisement entre hétérozygotes, ce qui induit une résolution limitée.

En revanche, pour les études d'associations, on va s'adresser potentiellement à une diversité génétique naturelle et élevée et une histoire de recombinaisons pouvant être très longue. Par conséquent le déséquilibre de liaison entre un locus fonctionnel et les marqueurs moléculaires est faible sauf pour ceux à très courte distance, ce qui donne une résolution beaucoup plus fine (Zhu *et al.*, 2008) (Figure 1.10).

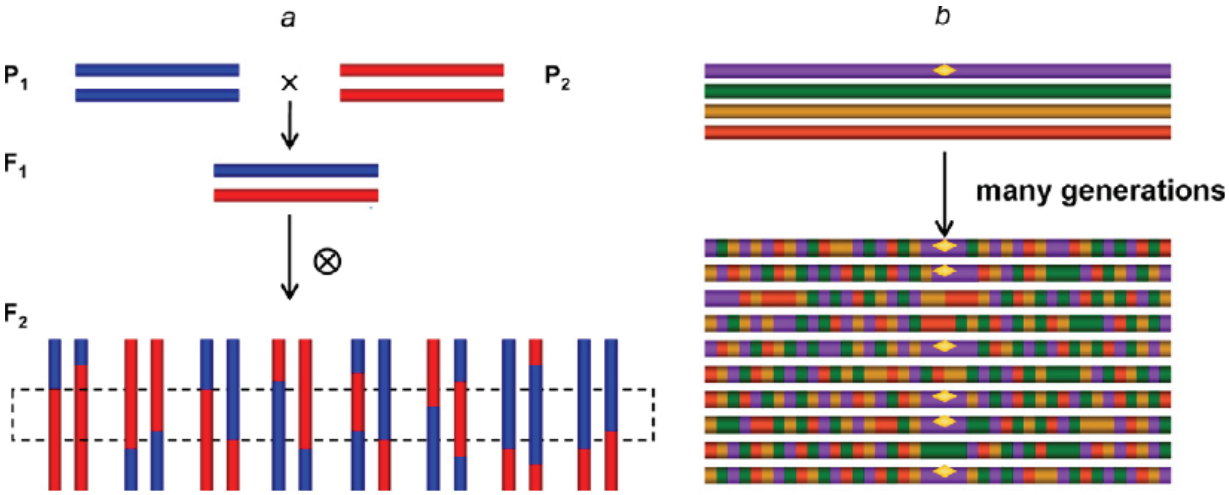


Figure 1.10 : Comparaison de l'étendue du DL (et donc de la résolution potentielle des associations marqueurs/caractères) entre les approches de cartographie classique (à gauche) et les études d'associations (à droite). Tiré de Zhu *et al.* (2008)

D'autre part le DL dans les populations de cartographie est maximisé et le DL du à la structure est nul, rendant la recherche d'associations plus simple et surtout le risque de détection de fausses associations moindre. Ces approches existent depuis plusieurs dizaines d'années et les modèles statistiques adaptés à celles-ci ont aujourd'hui une certaine maturité et sont particulièrement puissants. Néanmoins la mise en place de populations de cartographie peut être très longue, surtout pour des espèces pérennes. De plus ce type de population est développé dans un but unique.

En étude d'association, on espère pouvoir utiliser des populations naturelles ou d'amélioration déjà existantes et pouvant servir à d'autres études ou à des schémas de sélection. Toutes ces propriétés des études d'association expliquent les raisons pour lesquelles les recherches actuelles basées sur celles-ci connaissent un essor très important (Figure 1.11).

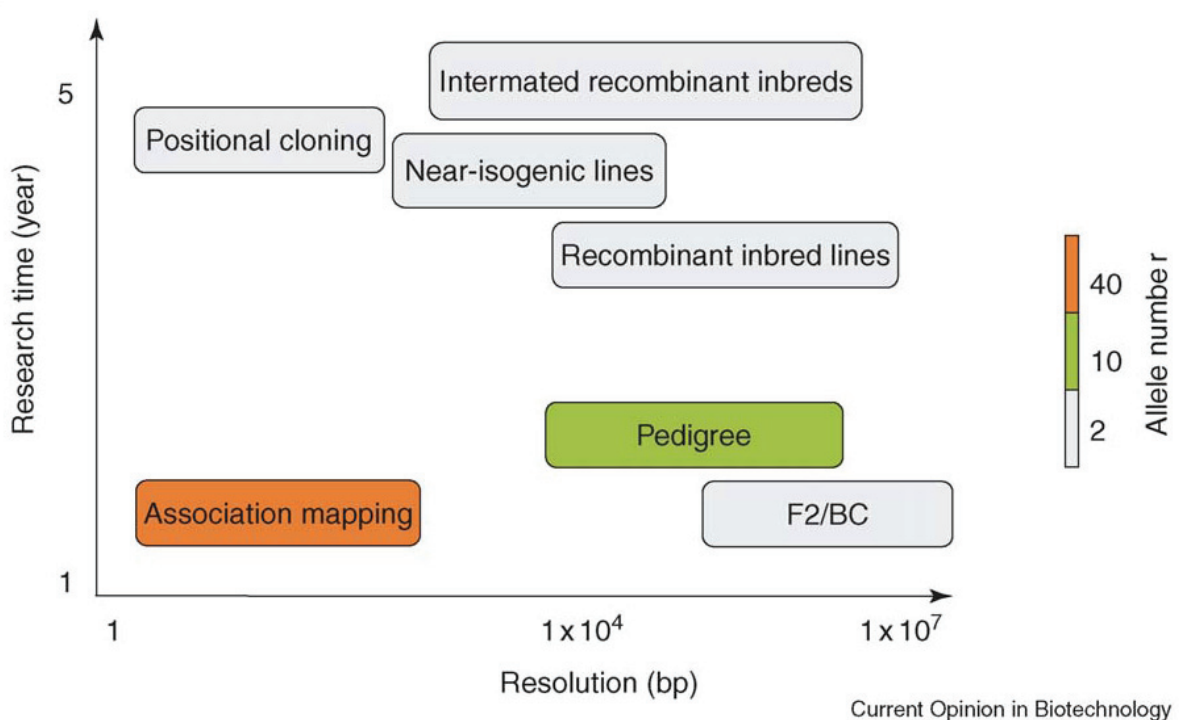


Figure 1.11 : Comparaison schématique de différentes méthodes d'identification d'associations entre polymorphismes moléculaires et variation phénotypique, en termes de résolution, de nombre d'allèles considérés et de temps de recherche. Tiré de Yu & Buckler (2006)

Quels modèles pour les études d'association ?

De nombreux modèles de génétique d'association pour la recherche d'associations entre marqueurs et caractères d'intérêt utilisant le déséquilibre de liaison ont été développés au cours des dernières années. Ces méthodes sont pour les plus importantes discutées dans l'article de Yu *et al.* (2006). Dans la plupart des cas, il apparaît qu'au minimum l'effet de la structure sur la détection d'association doit être pris en compte. Nous avons dans ce cas accès à trois modèles :

- Le modèle dit de Genomic Control proposé par Devlin & Roeder (1999)
- Le modèle de Structured Association proposé par Pritchard *et al.* (2000b)
- Le modèle comprenant à la fois l'effet de la structure et celui de l'apparentement (Q+K) proposé par Yu et al. (2006)

L'approche de Genomic Control (Devlin & Roeder, 1999; Devlin *et al.*, 2001) exploite le fait que la structure des populations génère une augmentation globale des

statistiques utilisées pour mettre en évidence des associations. En testant de multiples polymorphismes sur l'ensemble du génome, certains seulement étant liés au caractère d'intérêt, on peut estimer cette augmentation et la prendre en compte (Devlin *et al.*, 2001).

L'approche de Structured-Association avancée par Pritchard *et al.* (2000b) permet de s'affranchir plus efficacement de l'effet de la structure et des faux positifs impliqués par celle-ci. La méthode de Pritchard consiste à utiliser un ratio de vraisemblance (Λ) entre une hypothèse nulle dans laquelle les fréquences alléliques au locus candidat sont indépendantes du phénotype et une hypothèse alternative où les fréquences alléliques sont corrélées au phénotype. Cette méthode, initialement développée pour des études de type « case-control » a été adaptée pour des études d'association s'intéressant à des variations quantitatives par Thornsberry *et al.* (2001) : $\Lambda = \frac{\Pr_1(C;T,\hat{Q})}{\Pr_0(C;\hat{Q})}$

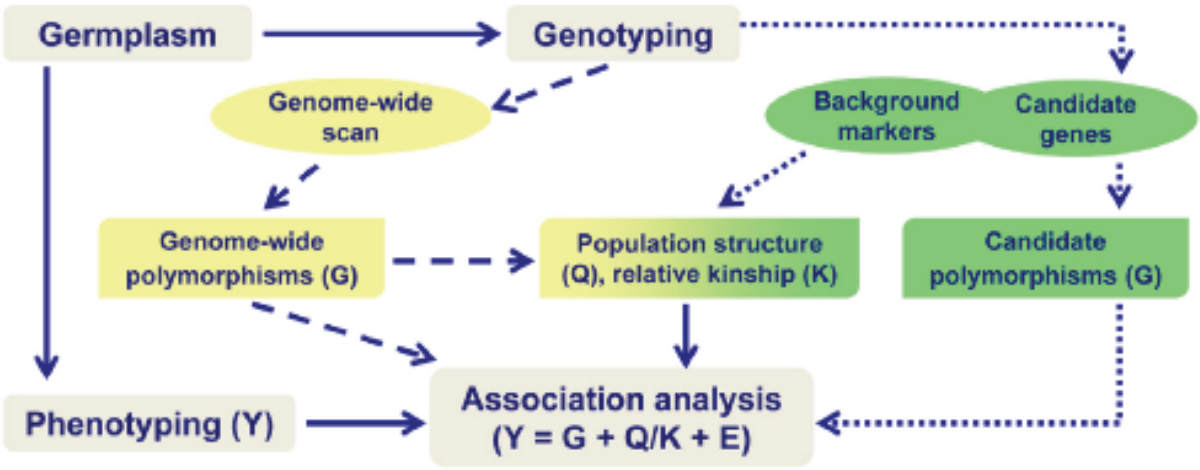
Dans cette équation C et T sont respectivement le génotype au polymorphisme candidat et la valeur du caractère pour chaque génotype, et Q la matrice de structure calculée à l'aide de Structure. Une régression logistique est utilisée pour calculer ces 2 probabilités avec pour variable cible le polymorphisme candidat et comme variables indépendantes T et Q. Un test de permutation est utilisé en parallèle pour estimer une p-value et tester la signification des associations détectées. Néanmoins cette approche est limitée à des modèles bialléliques.

Une approche similaire peut être menée en utilisant un modèle linéaire généralisé, cette fois-ci en prenant en compte les multiples allèles pouvant exister à un locus. Le modèle utilisé est alors $T = C + Q + \varepsilon$ ou T est la valeur du caractère, C le génotype au polymorphisme candidat, Q la matrice de structure et ε l'erreur résiduelle. Le seuil de signification est déterminé par permutations, une méthode moins conservative que la correction de Bonferroni (Flint-Garcia *et al.*, 2005). Cette méthode à permis de conduire avec succès des études d'association, notamment sur le maïs. Le gain de puissance de cette méthode par rapport aux méthodes de Genomic Control est indéniable, néanmoins pour certaines espèces dans lesquelles la structure génétique décrite par le modèle de Structure reste imprécise, notamment à cause d'apparentements multiples entre les individus, il semble qu'il y ait encore un taux de faux-positifs assez élevé ainsi qu'une perte de puissance considérable par rapport a ce que l'on pourrait obtenir avec un modèle plus adapté.

La prise en compte des apparentements dans les études d'association est rendu possible par l'introduction d'une matrice de kinship (K) comme variable dans un modèle mixte pour l'étude des corrélations marqueurs/caractère (Yu et al., 2006). Le modèle utilisé est alors $y = X\beta + S\alpha + Qv + Zu + e$ qui est une extension du modèle mixte traditionnel. Tous les effets fixes autres que les polymorphismes ou la structure des populations sont modélisés dans le terme Xβ. y est un vecteur des observations phénotypiques, β est un vecteur des effets fixes autres que les polymorphismes ou la structure des populations, α est un vecteur de l'effet des polymorphismes, v est un vecteur de l'effet de la population, u un vecteur de l'effet de l'apparentement, e un vecteur des effets résiduels. Q est une matrice obtenue de Structure reliant y à v. X, S et Z sont des matrices d'incidences composées de 1 et de 0 reliant respectivement y à β, α et u.

Il semble donc primordial de connaître au mieux la structure génétique de l'espèce et des populations afin de guider le choix du modèle à utiliser pour les études d'association. Enfin notons que la prise en compte de la structure et des apparentements comme variables dans les modèles utilisés peuvent mener à des faux négatifs, notamment par une correction trop importante des effets de structure génétique et d'apparentements. Ceci peut être observé dans le cas d'un caractère phénotypique dont la variation serait liée à la structure des populations, comme c'est par exemple le cas pour la précocité de floraison chez le maïs (Camus-Kulandaivelu, 2006; Camus-Kulandaivelu *et al.*, 2006).

Enfin, il est à souligner qu'il existe 2 grands types d'approches possibles en fonction de l'étendue du DL constaté dans les populations étudiées, les études au niveau génome entier ou les approches basées sur des régions ou des gènes candidats (Figure 1.12). Les différences entre ces 2 approches sont présentées dans la figure suivante. Par conséquent la connaissance de l'étendue du DL est primordiale pour le choix du type d'étude que l'on souhaite mener.



Genome-wide association mapping	Candidate-gene association mapping
It is a comprehensive approach to systematically search the genome for causal genetic variation. A large number of markers are tested for association with various complex traits, and prior information regarding candidate genes is not required. It works best for a research consortium with complementary expertise and adequate funding.	Candidate genes are selected based on prior knowledge from mutational analysis, biochemical pathway, or linkage analysis of the trait of interest. An independent set of random markers needs to be scored to infer genetic relationships. It is a low cost, hypothesis-driven, and trait-specific approach but will miss other unknown loci.

Figure 1.12 : Diagramme schématique et contraste entre les approches génome entier et gène candidats. Tiré de Zhu *et al.* (2008)

Démarche de la thèse et incorporation au programme de recherche sur le caféier

Nous avons vu que les études d'association pouvaient permettre un gain de temps important dans la recherche d'associations marqueurs/caractères. Cet aspect est particulièrement important pour les espèces pérennes telles que la nôtre. Les travaux de diversité et d'étude du déséquilibre de liaison au sein de notre espèce s'intègrent donc de manière évidente au programme de recherche sur l'amélioration de la qualité, de la productivité et de la tolérance à la sécheresse de *C. canephora*. Afin de proposer une stratégie d'étude d'association pour notre espèce, nous analyserons tout d'abord la structuration de la diversité génétique de *C. canephora*, préalable indispensable pour la mise en place d'une stratégie de Sélection Assistée par Marqueurs efficace et pour l'évaluation du DL. Cette étude de diversité se fera depuis le genre *Coffea* jusqu'à la population de base d'amélioration de *C. canephora*. Il sera ainsi possible de replacer la structuration de la diversité dans un ensemble

allant du genre à la plante. Il sera également possible de tester les possibilités d'extension des résultats obtenus sur *C. canephora* aux autres espèces.

Nous analyserons ensuite le DL à différents niveaux de structure. Ces travaux seront menés sur plusieurs populations avec différents types de marqueurs et à différentes échelles. Ceci afin de bien connaître le DL de nos populations et d'en tirer les éléments indispensables pour proposer des études d'association ciblées et adaptées à notre matériel. La démarche générale sur laquelle a reposé le travail de thèse est schématisée en Figure 1.13.

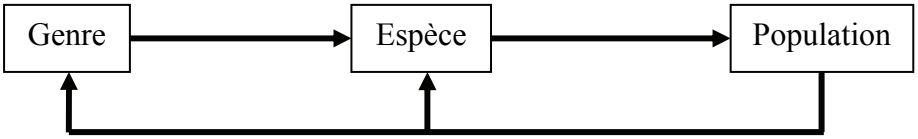


Figure 1.13 : Réflexion générale ayant guidé la démarche de la thèse

Ce premier chapitre était une introduction générale à la problématique de ce travail. Un second chapitre traitera de la diversité du genre *Coffea*, avec des implications pour la cartographie comparée et la généralisation des résultats obtenus sur *C. canephora* aux autres espèces du genre. Le troisième chapitre présentera une étude détaillée de la diversité intraspécifique de *C. canephora* et sa structuration. Il donne également quelques informations sur les résultats de cartographie génétique obtenus sur un croisement intergroupe. Le quatrième chapitre donnera une vue du DL au niveau du génome entier par microsatellites et les implications de celui-ci sur d'éventuelles études d'association. Le cinquième chapitre présentera, quant à lui, quelques résultats d'étude de DL à une échelle plus fine et une comparaison entre marqueurs microsatellites et polymorphismes de séquences.

Chapitre 2 : La diversité du genre *Coffea* par microsatellites

Introduction

Le genre *Coffea* présente une importance économique majeure pour les pays producteurs de café de la zone intertropicale, prenant la seconde place mondiale en termes de volume d'échanges économiques, juste derrière le pétrole. L'amélioration des caféiers pour une bonne valeur agronomique et la qualité des produits peut bénéficier aujourd'hui de méthodologies basées sur l'utilisation des marqueurs moléculaires

Le développement récent des études de cartographie comparée et de méta-analyses de QTL utilise des marqueurs moléculaires servant de « ponts » entre les différentes cartes génétiques considérées (Casasoli, 2004; Chagné, 2004; Veyrieras, 2006). Ces marqueurs ponts sont des marqueurs communs aux différentes cartes et permettent de relier celles-ci afin d'étudier plus précisément les QTLs, de les valider, de diminuer leur intervalles de confiance. L'enjeu de ces approches est important et nécessite d'avoir à disposition des marqueurs transférables non seulement au niveau de l'espèce mais aussi du genre (Yu *et al.*, 1999; Combes *et al.*, 2000; Alvarez *et al.*, 2001; Clauss *et al.*, 2002; Poncet *et al.*, 2004; Gao *et al.*, 2005). De plus, des approches comparatives sur plusieurs espèces proches ou éloignées permettent une réflexion sur la dynamique de mise en place du DL, y compris en considérant des espèces à régimes de reproduction contrastés ou ayant eu des histoires évolutives totalement différentes. Enfin ces différences entre les espèces pourraient amener à réaliser des études d'association de manière comparative entre plusieurs espèces afin de pouvoir répondre efficacement aux enjeux que représentent les recherches de corrélations entre marqueurs et caractères phénotypiques en validant celles-ci par leur répétabilité. Une autre potentialité de telles approches serait d'identifier dans des espèces voisines, et susceptibles de donner aisément des hybrides fertiles, des caractères intéressants pour l'amélioration qui pourraient alors être introgressés par sélection assistée par marqueurs.

L'étude qui suit s'inscrit dans le cadre d'une démarche visant à évaluer la faisabilité au niveau du genre *Coffea* de ce type d'approche comparative en utilisant des marqueurs aisés à mettre en œuvre et produisant de nombreuses données. Elle a pour but d'évaluer les possibilités d'extension des études que nous allons mener sur *C. canephora* aux autres

espèces du genre, et notamment *C. arabica*, *C. congensis* et *C. liberica* en testant l'amplification des marqueurs que nous utiliserons dans la suite de ce travail sur différentes espèces appartenant à l'ensemble de la diversité connue du genre.

Article publié dans *Genome*

Diversité des caféiers évaluée à l'aide de marqueurs SSR : structure du genre *Coffea* et perspectives pour l'amélioration.

Diversity in coffee assessed with SSR markers: structure of the genus *Coffea* and perspectives for breeding

Philippe Cubry, Pascal Musoli, Hyacinte Legnaté, David Pot, Fabien de Bellis, Valérie Poncet, François Anthony, Magali Dufour, and Thierry Leroy

Abstract: The present study shows transferability of microsatellite markers developed in the two cultivated coffee species (*Coffea arabica* L. and *C. canephora* Pierre ex Froehn.) to 15 species representing the previously identified main groups of the genus *Coffea*. Evaluation of the genetic diversity and available resources within *Coffea* and development of molecular markers transferable across species are important steps for breeding of the two cultivated species. We worked on 15 species with 60 microsatellite markers developed using different strategies (SSR-enriched libraries, BAC libraries, gene sequences). We focused our analysis on 4 species used for commercial or breeding purposes. Our results establish the high transferability of microsatellite markers within *Coffea*. We show the large amount of diversity available within wild species for breeding applications. Finally we discuss the consequences for future comparative mapping studies and breeding of the two cultivated species.

Key words: SSR markers, microsatellites, *Coffea*, transferability, cross-amplification, genetic diversity.

Résumé : La présente étude montre la transférabilité de marqueurs microsatellites développés sur les deux espèces de caféiers cultivées (*Coffea arabica* L. et *C. canephora* Pierre ex Froehn.) à 15 espèces représentant les principaux groupes précédemment identifiés du genre *Coffea*. L'évaluation de la diversité et des ressources génétiques disponibles au sein du genre *Coffea* et le développement de marqueurs moléculaires transférables d'une espèce à l'autre sont des étapes importantes pour l'amélioration de ces deux espèces. Nous avons travaillé sur 15 espèces avec 60 marqueurs microsatellites développés suivant différentes méthodologies (banques enrichies en microsatellites, banques BAC, séquences de gènes). Nous avons plus particulièrement analysé quatre espèces d'intérêt en commerce ou en amélioration. Nos résultats établissent que les microsatellites sont hautement transférables dans le genre *Coffea*. Nous mettons en évidence l'important réservoir de diversité pour l'amélioration que constituent les espèces sauvages de ce genre. Enfin nous discutons des implications pour de futures études de cartographie comparée et l'amélioration des deux espèces cultivées.

Mots-clés : marqueurs microsatellites, *Coffea*, transférabilité, amplification croisée, diversité génétique.

Introduction

The genus *Coffea* consists of 103 species (Davis and Stoffelen 2006) originated from intertropical regions of Africa and Madagascar. Only two species are cultivated: *C. arabica* L., which represents 65% of the world's coffee production, and *C. canephora* Pierre ex Froehn. Coffee species are diploid ($2n = 2x = 22$) except for *C. arabica*, which is tetraploid ($2n = 4x = 44$). *Coffea arabica* is self-compatible, like two diploid species, *C. heterocalyx* Stoff. and *C. anthonyi* Stoff. & F.Anthony (Davis and Stoffelen 2006). Previous phylogenetic studies based on other markers such as rDNA (Lashermes et al. 1997) and cpDNA variation (Cros et al. 1998) have shown that the genus *Coffea* is organized into 4 groups with different geographical origins, i.e., Central and West Africa (WC clade), East Africa (E clade), Central Africa (C clade), and Madagascar (M clade).

Microsatellite markers present different properties than the other markers previously used (such as RFLPs, isozymes, and cpDNA) and give a complementary view of the coffee genus diversity. SSR (simple sequence repeat) or microsatellite markers are highly variable and codominant

Received 5 April 2007. Accepted 10 October 2007. Published on the NRC Research Press Web site at genome.nrc.ca on 18 December 2007.

Corresponding Editor: F. Belzile.

P. Cubry,[1] **D. Pot, F. de Bellis, M. Dufour, and T. Leroy.** CIRAD, UMR DAP, TA A-96/03, avenue Agropolis, 34398 Montpellier CEDEX 5, France.
P. Musoli. Coffee Research Institute, P.O. Box 185, Mukono, Uganda.
H. Legnaté. CNRA, BP 808, DIVO, République de Côte d'Ivoire.
V. Poncet. IRD, UMR DIA-PC, 911 avenue Agropolis, BP 64501, 34394 Montpellier CEDEX 5, France.
F. Anthony. IRD, UMR RPB, 911 avenue Agropolis, BP 64501, 34394 Montpellier CEDEX 5, France.

[1]Corresponding author (e-mail: philippe.cubry@cirad.fr).

 doi:10.1139/G07-096

(Tautz and Renz 1984; Jarne and Lagoda 1996). They have already been analysed for their transferability within the coffee genus for 6 species, *C. canephora*, *C. eugenioides* S.Moore, *C. heterocalyx*, *C. liberica* Bull ex Hiern., *C. anthonyi*, and *C. pseudozanguebariae* Bridson (Poncet et al. 2004), and compared with AFLPs (Prakash et al. 2005). SSR markers have also been used to assess genetic diversity in the two main cultivated species, *C. arabica* and *C. canephora* (Anthony et al. 2002*a*, 2002*b*; Moncada and McCouch 2004; Cubry et al. 2005; Prakash et al. 2005). The present study gives cross-amplification results for a set of microsatellite markers in a larger sample of species and individuals.

In addition to a large survey of the transferability of the markers, we performed a detailed analysis of the two cultivated species (*C. arabica* and *C. canephora*) and two related species used for both quality and productivity improvement (*C. liberica* and *C. congensis*). A crisis of low prices has occurred during past years, and farmers have to produce a better quality coffee to maintain their incomes. Identifying the amount of genetic diversity available for improvement is especially important for *C. arabica*, which has been identified as a species with a very narrow genetic base (Anthony et al. 2002*a*). Since the genus *Coffea* diverged recently from others (5 to 25 million years ago; Lashermes et al. 1996), most of the species are genetically highly related and a lot of hybridizations are possible (Louarn 1992). Indeed, spontaneous and viable crosses of *C. canephora* × *C. congensis*, *C. arabica* × *C. liberica*, and *C. arabica* × *C. canephora* have been described (Cramer 1948; Prakash et al. 2002). These hybrids are widely used in breeding programs for resistance to pests and disease or for quality improvement.

In the present paper, we analyse the diversity of 15 *Coffea* species belonging to the 4 previously identified genetic groups using 60 microsatellite markers from different origins and covering the whole genome. We also detail the relationships among 4 species, 2 cultivated and 2 related wild ones. Finally, we discuss the consequences for breeding of *C. arabica* and *C. canephora*.

Materials and methods

Plant material

We used a total of 42 individuals from 15 *Coffea* species in our study (Table 1). Four species of particular interest were represented by more than 4 individuals to enable comparison of several diversity variables. These 4 species were *C. canephora*, *C. arabica*, *C. congensis*, and *C. liberica*.

For *C. arabica*, we studied both cultivated and wild accessions, including commercial hybrids between the two main cultivars, 'Typica' and 'Bourbon'. For *C. canephora* and *C. liberica*, we analysed, respectively, genotypes from different genetic groups (B, C, SG2, and Guinean) and varieties (*liberica*, *dewevrei*) chosen to represent the greatest diversity (Louarn 1992; Anthony 1992; Montagnon 2000; Dussert et al. 2003). *Coffea canephora* accessions also included new material from Uganda (Musoli et al. 2006), including wild material surveyed in Itwara Forest (UW) and the cultivar 'Nganda' (UN). *Coffea congensis* was represented by accessions from different Central African regions.

Eleven other species from different geographic origins covering the whole repartition of *Coffea* were included to provide an overview of the global diversity, including at least 2 species of each of the previously described diversity clades (i.e., C, WC, E, and M).

Coffea canephora genotypes were kindly provided by the CNRA (Centre National de Recherche Agronomique) from field collection in Divo (République de Côte d'Ivoire). Wild *C. canephora* (UW) and 'Nganda' (UN) genotypes from Uganda were conveniently provided by the CORI (Coffee Research Institute) of Uganda. *Coffea arabica*, *C. congensis*, *C. liberica*, and *C. sessiliflora* Bridson genotypes came from field collections in French Guiana. One individual of each of these 4 species was kindly provided by the IRD (Institut de Recherche pour le Développement) greenhouse collection in Montpellier, France. Material of 9 other species also came from the IRD collection: *C. anthonyi*, previously known as *C. 'sp. Moloundou'*, *C. bertrandii* A.Chev., *C. eugenioides*, *C. humilis* A.Chev., *C. millotii* J.-F.Leroy, *C. pseudozanguebariae*, *C. racemosa* Lour., *C. salvatrix* Swynn. & Philipson, and *C. stenophylla* G.Don.

DNA extraction

Genomic DNA was extracted from ground leaves following an extraction method using a MATAB buffer adapted from Risterucci et al. (2000). A purification of the extracts using products from the solution-based Wizard® SV Genomic DNA Purification System (Promega, Madison, Wisconsin, USA, Cat. No. A1125) was then performed.

Microsatellite markers

In this study, we used microsatellite markers obtained from different origins (Table 2). DLxxx primers were previously published and developed from a *C. canephora* BAC library (Leroy et al. 2005). A second set came from a microsatellite motif–enriched library of *C. canephora* clone 126 (Dufour et al. 2001) and from an enriched library of *C. arabica* 'Caturra' (Rovelli et al. 2000). Primers for the enriched *C. arabica* library came from Poncet et al. (2004) and primers for the enriched *C. canephora* library were designed by Poncet et al. (2007) using Primer3 software (Rozen and Skaletski 2000). SSRxxx primers were designed from sequences of sucrose synthase (*SuSy*) genes (Geromel et al. 2006) using Primer3 (D. Pot, unpublished data). A total of 60 loci were screened in this study and all of them, except SSRxxx loci, have been mapped on an intraspecific *C. canephora* genetic map (T. Leroy, unpublished data).

PCR and data acquisition

For each reaction, 2.5 ng of DNA template was mixed with 5 μL of PCR buffer (10 mmol/L Tris-HCl, 50 mmol/L KCl, 2 mmol/L $MgCl_2$, 0.001% glycerol), 200 μmol/L dNTPs, 0.10 μmol/L of reverse primer, 0.08 μmol/L of forward primer tailed with M13 sequence, 0.10 μmol/L of fluorescently labelled M13 primer, and 0.1 U of *Taq* DNA polymerase. PCR amplifications were performed in an Eppendorf Mastercycler ep 384 (Eppendorf, Westbury, New York, USA). The amplification program consisted of an initial denaturation cycle of 4 min at 94 °C followed by 9 cycles of "touch-down" PCR consisting of 45 s at 94 °C, 1 min at 60 °C to 55 °C, decreasing by 0.5 °C each cycle, and 1 min 30 s at 72 °C. The next 26 cycles

Table 1. List of plant material and providers.

Coffea species	Working name	Variety or diversity group	Collection
Species of particular interest for commercial or breeding purposes			
C. arabica	Arabica_1	'Caturra'	IRD, France
C. arabica	Arabica_2	'Red Catuaí 1'	CIRAD, French Guiana
C. arabica	Arabica_3	'Guinee pita 1'	CIRAD, French Guiana
C. arabica	Arabica_4	'Sidamo 1'	CIRAD, French Guiana
C. arabica	Arabica_5	'Mundo Novo'	CIRAD, French Guiana
C. arabica	Arabica_et1	Wild ethiopian	CIRAD, French Guiana
C. arabica	Arabica_et2	Wild ethiopian	CIRAD, French Guiana
C. arabica	Arabica_et3	Wild ethiopian	CIRAD, French Guiana
C. canephora	Can_b1	Congolese group B	CNRA, République de Côte d'Ivoire
C. canephora	Can_c1	Congolese group C	CNRA, République de Côte d'Ivoire
C. canephora	Can_sg2_1	Congolese group SG2	CNRA, République de Côte d'Ivoire
C. canephora	Can_g1	Guinean	CNRA, République de Côte d'Ivoire
C. canephora	Can_g2	Guinean	CNRA, République de Côte d'Ivoire
C. canephora	Can_u1	Uganda, 'Nganda'	CORI, Uganda
C. canephora	Can_u2	Uganda, wild	CORI, Uganda
C. canephora	Can_u3	Uganda, wild	CORI, Uganda
C. canephora	Can_g3	Guinean	CNRA, République de Côte d'Ivoire
C. congensis	Congensis_1		IRD, France
C. congensis	Congensis_2		CIRAD, French Guiana
C. congensis	Congensis_3		CIRAD, French Guiana
C. congensis	Congensis_4		CIRAD, French Guiana
C. congensis	Congensis_5		CIRAD, French Guiana
C. liberica	Liberica_1		IRD, France
C. liberica	Liberica_2_1	*liberica*	CIRAD, French Guiana
C. liberica	Liberica_3_1	*liberica*	CIRAD, French Guiana
C. liberica	Liberica_4_1	*liberica*	CIRAD, French Guiana
C. liberica	Liberica_5_d	*dewevrei*	CIRAD, French Guiana
C. liberica	Liberica_6_d	*dewevrei*	CIRAD, French Guiana
C. liberica	Liberica_7_d	*dewevrei*	CIRAD, French Guiana
Other species included in this study			
C. anthonyi	Anthonyi		IRD, France
C. bertrandii	Bertrandii		IRD, France
C. brevipes	Brevipes		IRD, France
C. eugenioides	Eugenioides		IRD, France
C. humilis	Humilis		IRD, France
C. milloti	Milloti		IRD, France
C. pseudozanguebariae	Pseudozanguebariae		IRD, France
C. racemosa	Racemosa		IRD, France
C. salvatrix	Salvatrix		IRD, France
C. sessiliflora	Sessiliflora_1		IRD, France
C. sessiliflora	Sessiliflora_2		CIRAD, French Guiana
C. sessiliflora	Sessiliflora_3		CIRAD, French Guiana
C. stenophylla	Stenophylla		IRD

consisted of 94 °C for 45 s, 55 °C for 1 min, and 72 °C for 1 min 30 s prior to a final elongation step at 72 °C for 5 min.

Fluorescently labelled PCR products were analysed by electrophoresis on a 6.5% polyacrylamide gel using a LI-COR® 4300 automated sequencer (LI-COR Biosciences, Lincoln, Nebraska, USA). Gel images were retrieved and annotated with the manufacturer's program SAGA^{GT}. We assigned allele sizes manually to each individual on the basis of the automated analyses of SAGA^{GT}. Previously studied individuals of *C. canephora* (Cubry et al. 2005) served as controls. The data matrix was exported as a text file and formatted in Excel® software for the different programs used for the analysis.

Data analysis

A dissimilarity matrix was computed from the data file using the software DARwin 5 (Perrier et al. 2003). The dissimilarities were calculated using a simple matching distance index. Since *C. arabica* exhibited a maximum of 2 alleles per locus in our data, we decided to manage genotypes from this species as diploid genotypes. The dissimilarity matrix was used to infer a global diversity tree using the weighted neighbor-joining method (Saitou and Nei 1987) as

Table 2. List of the 60 SSR markers used in the study.

EMBL acc. No.	Marker name	Repeat type	No. of repeats	Primer sequences (5′→3′)	Sequence origin	Primer origin	Species of origin
AJ250257	257	CA	9	F: GACCATTACATTTCACACAC R: GCATTTTGTTGCACACTGTA	Combes 2000	Poncet 2004	*Coffea arabica* 'Caturra'
AM231186	305	TG	8	F: AACTTCACTAATCTGTTGTTGCTG R: GCACATCTATCCATCTTTTGG	Dufour 2001	Poncet 2007	*Coffea canephora*, clone 126
AM231546	327	CA	9	F: GGCTCAAAATCACCCTTTGT R: CTAGGATCGTGGCAGAAGAAG	Dufour 2001	Poncet 2007	*Coffea canephora*, clone 126
AM231547	329	GT	10	F: ACTCAGACAAACCCTTCAAC R: GATGTTTTGCATCTATTTGG	Dufour 2001	Poncet 2007	*Coffea canephora*, clone 126
AM231548	334	AC	8	F: TATGCCTCAGCACCTATCTA R: TACTTCCCCTGTTCCTTATG	Dufour 2001	Poncet 2007	*Coffea canephora*, clone 126
AM231549	341	CA, TA	12, 5	F: CATTGGTGTCAAGGGTCAAG R: AAAGTATCAGAAGGAAAAGTCTCGTAA	Dufour 2001	Poncet 2007	*Coffea canephora*, clone 126
AM231550	350	GT	8	F: TCAAAAGAGGGCACGAA R: ACGACAATAACTTTGCATGTCT	Dufour 2001	Poncet 2007	*Coffea canephora*, clone 126
AM231551	351	GT	13	F: AAGGATGGCAAGTGGATTTCT R: GCAGCTCTTGATTGTAGTTTCGT	Dufour 2001	Poncet 2007	*Coffea canephora*, clone 126
AM231552	355	TG	15	F: CTATGATGTCTTCCAACCTTCTAAC R: GGTCCAATTCTGTTTCAATTTC	Dufour 2001	Poncet 2007	*Coffea canephora*, clone 126
AM231553	356	TG	14	F: TGAAGTCAACCTGAATACCAGA R: ACGCACGCACGAATG	Dufour 2001	Poncet 2007	*Coffea canephora*, clone 126
AM231554	358	CA	11	F: CATGCACTATTATGTTTGTGTTTT R: TCTCGTCATATTTACAGGTAGGTT	Dufour 2001	Poncet 2007	*Coffea canephora*, clone 126
AM231555	360	CA	10	F: ACAGTAGTATTTCATGCCACATCC R: ACATTTGATTGCCTCTTGACC	Dufour 2001	Poncet 2007	*Coffea canephora*, clone 126
AM231556	364	A	21	F: AGAAGAATGAAGACGAAACACA R: TAACGCCTGCCATCG	Dufour 2001	Poncet 2007	*Coffea canephora*, clone 126
AM231557	367	AC	12	F: TCAATCCCTGTATTCCTGTTT R: CTAGGCACTTAAAATCTCTATAACG	Dufour 2001	Poncet 2007	*Coffea canephora*, clone 126
AM231558	368	TG	13	F: CACATCTCCATCCATAACCATTT R: TCCTACCTACTTGCCTGTGCT	Dufour 2001	Poncet 2007	*Coffea canephora*, clone 126
AM231559	371	CA	9	F: AGACACACAAGGCAATAATCAAAC R: TCTTGAGCAGCATGGGAAC	Dufour 2001	Poncet 2007	*Coffea canephora*, clone 126
AM231560	384	AC	10	F: ACGCTATGACAAGGCAATGA R: TGCAGTAGTTTCACCCTTTATCC	Dufour 2001	Poncet 2007	*Coffea canephora*, clone 126
AM231561	388	CA	9	F: ATGAAACGAGAATCCATACCCTAC R: AGAGGTAAAAGGAAAATGCTAGACC	Dufour 2001	Poncet 2007	*Coffea canephora*, clone 126
AM231562	392	TC	16	F: AAGGTATTGGTCTGCCTTTGT R: CTAACCCTAATCCCCAGCA	Dufour 2001	Poncet 2007	*Coffea canephora*, clone 126
AM231563	394	TG	9	F: GCCGTCTCGTATCCCTCA R: GAAGCCAGAAAGTCAGTCACATAG	Dufour 2001	Poncet 2007	*Coffea canephora*, clone 126
AM231564	395	GT	13	F: CATCATTTTGTTGGCAAAG R: TGGTTATTTCCTTCTTTGTATTG	Dufour 2001	Poncet 2007	*Coffea canephora*, clone 126

Table 2 (*continued*).

EMBL acc. No.	Marker name	Repeat type	No. of repeats	Primer sequences (5′→3′)	Sequence origin	Primer origin	Species of origin
AM231565	429	A	13	F: CATTCGATGCCAACAACCT R: GGGTCAACGCTTCTCCTG	Dufour 2001	Poncet 2007	*Coffea canephora*, clone 126
AM231566	442	CA	19	F: CGCAAATCTGAGTATCCCAAC R: TGGATCAACACTGCCCTTC	Dufour 2001	Poncet 2007	*Coffea canephora*, clone 126
AM231567	445	AC	10	F: CCACAGCTTGAATGACCAGA R: AATTGACCAAGTAATCACCGACT	Dufour 2001	Poncet 2007	*Coffea canephora*, clone 126
AM231568	456	AC	14	F: TGGTTGTTTTCTTCCATCAATC R: TCCAGTTTCCCACGCTCT	Dufour 2001	Poncet 2007	*Coffea canephora*, clone 126
AM231569	460	CA	11	F: TGCCTTCAAAATGCTCTATAACC R: GCTGATATTCTTGGATGGAGTTG	Dufour 2001	Poncet 2007	*Coffea canephora*, clone 126
AM231570	461	AC	9	F: CGGCTGTGACTGATGTG R: AATTGCTAAGGGTCGAGAA	Dufour 2001	Poncet 2007	*Coffea canephora*, clone 126
AM231571	463	AC	8	F: CATTCTTCCCACGATTCTATCTC R: GTGACTTTCGGTTGAAATACTGG	Dufour 2001	Poncet 2007	*Coffea canephora*, clone 126
AM231572	471	CT	12	F: TTACCTCCCGGCCAGAC R: CAGGAGACCAAGACCTTAGCA	Dufour 2001	Poncet 2007	*Coffea canephora*, clone 126
AM231573	472	CA, TA	8, 8	F: AATCATGGGGACAGGACAAG R: TCTGCTAGACTTGACATCTTTTGG	Dufour 2001	Poncet 2007	*Coffea canephora*, clone 126
AM231574	477	AC	16	F: CGAGGGTTGGGAAAAGGT R: ACCACCTGATGTTCCATTTGT	Dufour 2001	Poncet 2007	*Coffea canephora*, clone 126
AM231575	495	AC	8	F: CATGGATGGGAAGGCAGT R: CTTGGAAAACTTGCTGAATGTG	Dufour 2001	Poncet 2007	*Coffea canephora*, clone 126
AM231576	501	TG	8	F: CACCACCATCTAATGCACCT R: CTGCACCAGCTAATTCAAGC	Dufour 2001	Poncet 2007	*Coffea canephora*, clone 126
AJ308753	753	CA	15	F: GGAGACGCAGGTGGTAGAAG R: TCGAGAAGTCTTGGGGTGTT	Rovelli 2000	Poncet 2004	*Coffea arabica* 'Caturra'
AJ308755	755	CA	20	F: CCCTCCCTCTTTCTCCTCTC R: TCTGGGTTTTCTGTGTTCTCG	Rovelli 2000	Poncet 2004	*Coffea arabica* 'Caturra'
AJ308774	774	CT, CA	5, 7	F: GCCACAAGTTTCGTGCTTTT R: GGGTGTCGGTGTAGGTGTATG	Rovelli 2000	Poncet 2004	*Coffea arabica* 'Caturra'
AJ308779	779	TG	17	F: TCCCCCATCTTTTTCTTTCC R: GGGAGTGTTTTTGTGTTGCTT	Rovelli 2000	Poncet 2004	*Coffea arabica* 'Caturra'
AJ308782	782	GT	15	F: AAAGGAAAATTGTTGGCTCTGA R: TCCACATACATTTCCCAGCA	Rovelli 2000	Poncet 2004	*Coffea arabica* 'Caturra'
AJ308790	790	GT	21	F: TTTTCTGGGTTTTCTGTGTTCTC R: TAACTCTCCATTCCCGCATT	Rovelli 2000	Poncet 2004	*Coffea arabica* 'Caturra'
AJ308809	809	TGA	11	F: AGCAAGTGGAGCAGAAGAAG R: CGGTGAATAAGTCGCAGTC	Rovelli 2000	Poncet 2004	*Coffea arabica* 'Caturra'
AJ308837	837	TG, GA	16, 11	F: CTCGCTTTCACGCTCTCTCT R: CGGTATGTTCCTCGTTCCTC	Rovelli 2000	Poncet 2004	*Coffea arabica* 'Caturra'
AJ308838	838	AC	9	F: CCCGTTGCCATCCTTACTTA R: ATACCCGATACATTTGGATACTCG	Rovelli 2000	Poncet 2004	*Coffea arabica* 'Caturra'

Table 2 (*concluded*).

EMBL acc. No.	Marker name	Repeat type	No. of repeats	Primer sequences (5′→3′)	Sequence origin	Primer origin	Species of origin
AJ871882	DL003	CAAT	5	F: TAACAGAAGCACCAAAACC R: TCTAAACCCACCTCACAAC	Leroy 2005	Leroy 2005	*Coffea canephora*, clone 126
AJ871889	DL010	A	14	F: TAGTCCCTTTTCAGTGGT R: TTTCTTTGTTACGGAGTG	Leroy 2005	Leroy 2005	*Coffea canephora*, clone 126
AJ871890	DL011	GCT, CAT	4, 8	F: ATACATAAGCAAGCACTGA R: CAGAACAAATGAAATGGA	Leroy 2005	Leroy 2005	*Coffea canephora*, clone 126
AJ871892	DL013	CA, CT	6, 8	F: AGAGGGATGTCAGCATAA R: ATTTGTGTTTGGTAGATGTG	Leroy 2005	Leroy 2005	*Coffea canephora*, clone 126
AJ871899	DL020	T	23	F: TGCTCAAACTTCTTGCT R: CGCCAACTCTAATGTGT	Leroy 2005	Leroy 2005	*Coffea canephora*, clone 126
AJ871904	DL025	C	17	F: TTGTTGAGAGTGGAGGA R: CCAAAGACAGTGCAGTAA	Leroy 2005	Leroy 2005	*Coffea canephora*, clone 126
AJ871905	DL026	A	17	F: CGAGACGAGCATAAGAA R: GCTGGAATGAAGAATGTAG	Leroy 2005	Leroy 2005	*Coffea canephora*, clone 126
AJ871911	DL032	TACG	3	F: TGTTGGTGAAGAAATCC R: ATGGAGACAGGAAATAAAC	Leroy 2005	Leroy 2005	*Coffea canephora*, clone 126
AM231577	SSR001	T	3	F: CAATACGGCATGCATTTGAC R: TGTTGAACACGCAATTGACC	Geromel 2006	Pot 2006	*Coffea canephora*, clone 126
AM231578	SSR003	A	6	F: ATTTGCGTGCTGGATGTTTT R: ACCATGTAGGAAGGCCACAG	Geromel 2006	Pot 2006	*Coffea canephora*, clone 126
AM231579	SSR004	T	9	F: CCAACCCTAAGATGATTTTTGT R: AACCCCTCTCAAAACCCAGT	Geromel 2006	Pot 2006	*Coffea canephora*, clone 126
AM231582	SSR005	GAT	2	F: ATGTGGTGCTGATGTGCAGT R: GTCACGTGGGATGATGAGAA	Geromel 2006	Pot 2006	*Coffea canephora*, clone 126
AM231580	SSR009	GAAAA	5	F: CAAACAAAACAGTACAATTCAATCC R: ATCCCTGCGAGACCTGACTA	Geromel 2006	Pot 2006	*Coffea canephora*, clone 126
AM231581	SSR010	ATT	2	F: CGAAAGGAACACAGGAACCA R: CAGTGGTGAACTTAATCGTCCA	Geromel 2006	Pot 2006	*Coffea canephora*, clone 126
AM231583	SSR014	T	14	F: GGATCTTATCGCAATGAACCA R: CCAACAGTGTCCTTGCTGAA	Geromel 2006	Pot 2006	*Coffea canephora*, clone 126
AM231584	SSR015	T	12	F: TTCTTCACAAGAACCAACCCTAA R: AACCCCTCTCAAAACCCAAT	Geromel 2006	Pot 2006	*Coffea canephora*, clone 126
AM231585	SSR016	T	13	F: TGGTCAATTTGAAGCGACTG R: CCTCCATCCTTTCCCTTACC	Geromel 2006	Pot 2006	*Coffea canephora*, clone 126
AM231586	SSR017	TA	7	F: TGTTCCTCTGGCTGTTGATG R: CCGGTTGAATGAGGGTAAAG	Geromel 2006	Pot 2006	*Coffea canephora*, clone 126

implemented in DARwin. Five thousand bootstrap iterations were calculated to test the robustness of the nodes. Considering that some species were represented by more than one individual, we inferred another diversity tree with one randomly chosen individual per species. This tree allowed a better understanding of the genetic relationships between species without the interference of sampling size per species. The same inference method used for the global tree was used for this second tree.

Several genetic variables (e.g., number of alleles, gene diversity, and observed heterozygosity) were calculated using PowerMarker software (Liu and Muse 2005) for the global sample and for each of the 4 species of particular interest. We also computed the percentage of polymorphic loci by species. Ninety-five percent confidence intervals for each variable were estimated by performing 5000 bootstrap iterations across loci.

Results

Amplifications across the genus

The availability (percentage of amplification) per marker ranged from 30.9% to 100% among the 42 analysed genotypes, with a mean of 81.5% calculated from the raw matrix of observations (see Table S1[2]). Even if 3 markers appeared to be specific to the Central Africa clade, good transferability of microsatellites across *Coffea* species was observed.

The percentage of amplification per individual ranged from 51.7% for one *C. liberica* genotype (note that the mean for all *C. liberica* species is about 72%) to 98.3% for one *C. canephora* genotype. Values obtained here are close to those found by Poncet et al. (2004). For the 4 main species, amplification ranged from 72% for *C. liberica* to 89% for *C. arabica* and 90% for *C. canephora*. Amplification for *C. congensis* was intermediate (83%).

Genus diversity analysis

Figure 1 presents the neighbor-joining tree for the 42 individuals of the study based on 60 microsatellite loci. Bootstrap values greater than 40 are shown; this threshold was arbitrarily chosen for the readability of the figure. Ten diversity groups were discriminated by the analysis. The 4 genetic groups WC, C, E, and M, previously described by Lashermes et al. (1997), are indicated on this figure.

Groups C, E, and M were discriminated by our study, whereas species of the WC clade were classified in 7 different groups. *Coffea arabica* and *C. congensis* constituted original groups, while *C. canephora* and *C. liberica* were each represented by two groups. These two groups correspond to different geographical origins (Central and West Africa), as previously described by Berthaud (1986). For *C. liberica*, these two groups appear to be the varieties, *C. liberica* var. *liberica* and *C. liberica* var. *dewevrei*. For *C. canephora* the two groups correspond to the Guinean (G) clade and the Congolese clade, including the B and SG2 diversity groups. We observed strong relationships between B, SG2, and related Ugandan accessions (UW, UN), as previously described (Musoli et al. 2006). *Coffea brevipes* can be grouped with the Central African (Congolese) clade of *C. canephora*, while *C. humilis* and *C. stenophylla* appear to be grouped.

Within *C. arabica*, wild and cultivated materials were differentiated, as expected from previous studies of a small number of SSR markers (Anthony et al. 2002*a*, 2002*b*). The cultivated varieties represent a narrow genetic base, since dissimilarity distances between those genotypes are the shortest of the dendrogram.

The second tree, considering only one individual per species, allows us to describe 5 different groups for our sampled species. Groups M, C, and E are still discriminated, while species from West Africa (WC clade) are separated into two groups: *C. arabica*, *C. canephora*, and the related species *C. congensis* and *C. brevipes* form one group, while *C. liberica*, *C. humilis*, and *C. stenophylla* form another group. Bootstrap values supporting these groups are quite high for microsatellite markers.

The global diversity is high, with a mean gene diversity of 0.72 ± 0.03 and a mean allele number of 10.8 (see Table 3 for details). The number of alleles varies from 1 to 22 according to the locus considered. Of the total number of alleles (648), 304 (47%) are specific to one species. A complete table of private alleles is given as supplementary material (Table S2[2]). The percentage of the total number of private alleles for each species ranges from 0% for *C. anthonyi* to 31.25% for *C. canephora*, with a mean of 6.45% (see Table S3[2]). These results show the great amount of interspecific diversity within the genus, even if some species are represented by only one individual.

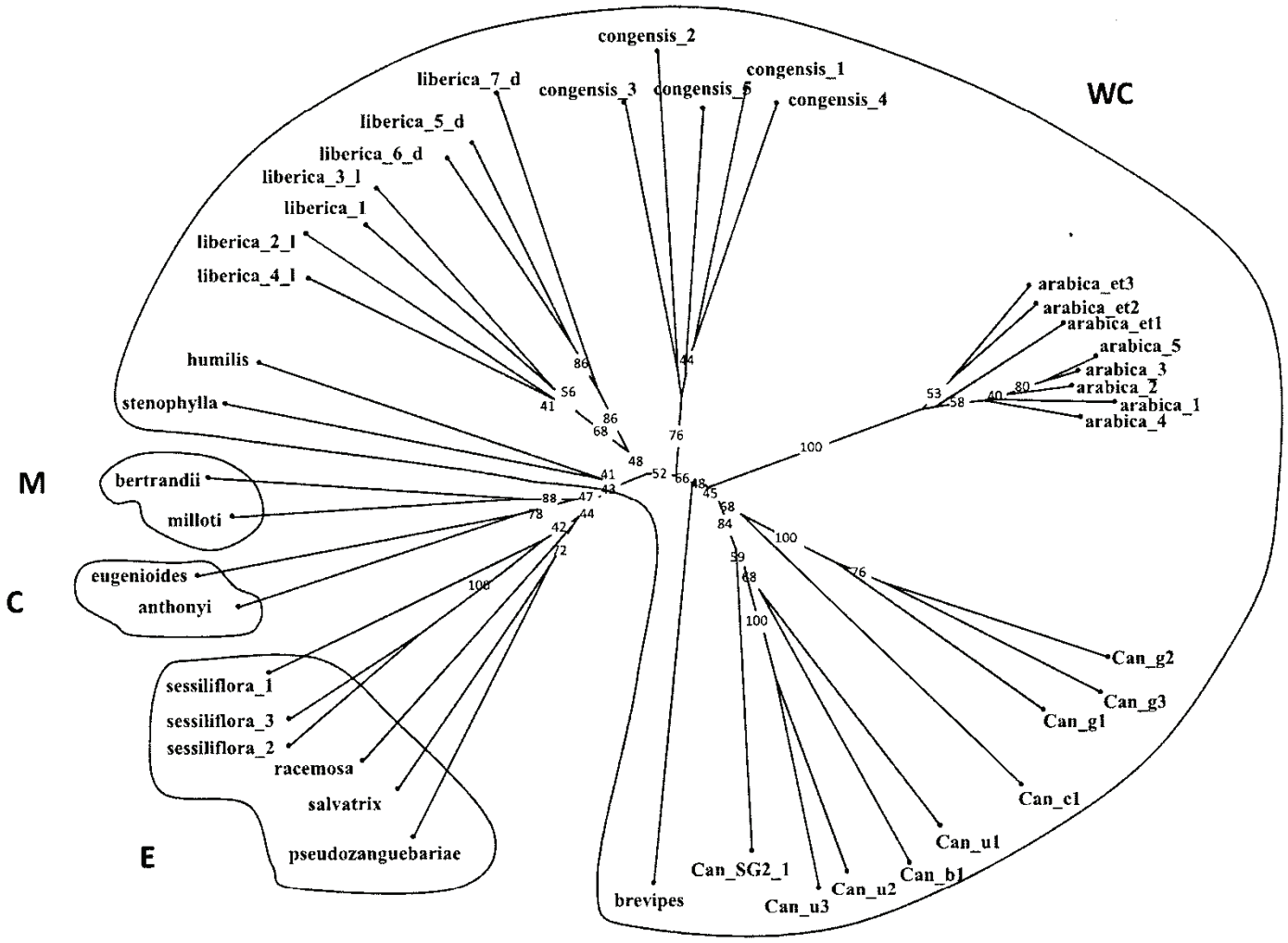
Considering the global sample, 59 markers are polymorphic. Only one, SSR016, which derived from a genic sequence, exhibited no polymorphism. At the intraspecific level, 91.7%, 75%, 76.7%, and 65% of the markers are polymorphic in *C. canephora*, *C. congensis*, *C. liberica*, and *C. arabica*, respectively. For the other species, polymorphism information should not be taken into consideration because only one or a small number of individuals are available.

Diversity analysis of several species

Four species of particular interest because of their economic importance or breeding potential were more accurately analysed in our study. This subsample of 4 species contributed an important part of the global sample diversity, with a mean number of alleles of 8. On the species diversity diagram (Fig. 2) they appear to be in 2 related clades. Table 3 presents the results for allele number, gene diversity, and observed heterozygosity for *C. arabica*, *C. canephora*, *C. congensis*, and *C. liberica* (Table S4[2] presents values calculated for all the species). *Coffea arabica* shows the lowest diversity, with a mean number of alleles of 2.10. Moreover, it is the only species that shows gene diversity less than observed heterozygosity. The global amount of diversity in these 4 species is important, with a mean gene diversity higher than 0.35. *Coffea canephora* appears to be the most

[2] Supplementary data for this article are available on the journal Web site (http://genome.nrc.ca) or may be purchased from the Depository of Unpublished Data, Document Delivery, CISTI, National Research Council Canada, Building M-55, 1200 Montreal Road, Ottawa, ON K1A 0R6, Canada. DUD 5250. For more information on obtaining material refer to http://cisti-icist.nrc-cnrc.gc.ca/irm/unpub_e.shtml.

Fig. 1. Neighbor-joining tree for the 42 individuals analyzed based on the dissimilarity matrix calculated by simple matching. Bootstrap values were calculated with 5000 repetitions; only values greater than or equal to 40 are shown.



diverse, with a gene diversity of 0.55 and a mean number of alleles of 5.00.

Discussion

Coffea diversity

The global amount of diversity within *Coffea* appears to be high. Considering the 4 previously described clades, we show that 3 groups can be confirmed (i.e., groups C, M, and E), while the fourth (WC) appears divided in two (Fig. 2). This division can be imputed to the use of SSRs, which have different properties than the previously used markers, and the high number of markers used in this study compared with the previous studies. Indeed, the high rate of mutation for microsatellite markers helps us to better investigate structure within species and species complexes.

Moreover, microsatellites are valuable tools to assess genetic structure at the species level, as demonstrated by the global diversity diagram (Fig. 1). This figure shows the relationships within 4 species of the WC clade, indicating structure at the intraspecific level for *C. liberica*, *C. canephora*, and *C. arabica*. In contrast, *C. congensis* appears to be homogeneous, at least for the genotypes studied.

Finally, we validated our sampling strategy, which consisted of analysing at least 2 species per previously known diversity clade for the whole genus to have an overview of the global genus diversity. We sampled more genotypes for 4 species particularly well known and of important economic and breeding interest (Lashermes et al. 1997; Anthony 1992; Poncet et al. 2004).

Our results validate the microsatellite-based approach to quickly study *Coffea* species by covering the entire genome, while sequence-based studies are generally limited to small numbers of genomic regions.

Transferability of microsatellite markers

We have confirmed the transferability of SSR markers across the genus *Coffea* for a larger sample of species than previously described. SSRs are useful markers for comparative studies across genera (Casasoli 2004). Their transferability over species across a genus has been shown for several genera including *Lycopersicon* (Alvarez et al. 2001), *Oryza* (Gao et al. 2005), *Vigna* (Yu et al. 1999), and *Coffea* (Combes et al. 2000; Poncet et al. 2004). Newly developed microsatellites based on *C. canephora* sequences exhibit the same properties as those previously developed based on

Table 3. Summary statistics calculated for the 60 SSR markers for the global sample (all 15 species studied), the 4 species focused on, and each of the 4 species separately.

	15 species			*C. arabica*			*C. canephora*			*C. congensis*			*C. liberica*			4 species		
Marker	*N*	GD	H_o	*N*	GD	H_o	*N*	GD	H_o	*N*	GD	H_o	*N*	GD	H_o	*N*	GD	H_o
DL003	6	0.61	0.37	2	0.23	0.29	2	0.40	0.14	1	0.00	0.00	3	0.37	0.50	3	0.57	0.25
DL010	12	0.77	0.36	2	0.19	0.22	5	0.57	0.11	4	0.54	0.80	3	0.56	1.00	8	0.72	0.45
DL011	7	0.62	0.17	1	0.00	0.00	4	0.61	0.22	3	0.29	0.20	3	0.47	0.14	6	0.63	0.16
DL013	12	0.86	0.30	2	0.50	1.00	3	0.45	0.00	1	0.00	0.00	3	0.51	0.00	7	0.83	0.38
DL020	13	0.86	0.43	4	0.56	0.89	6	0.69	0.50	2	0.26	0.00	5	0.64	0.43	11	0.86	0.50
DL025	7	0.78	0.34	3	0.54	1.00	3	0.53	0.11	3	0.29	0.20	3	0.51	0.00	6	0.74	0.37
DL026	13	0.81	0.11	1	0.00	0.00	5	0.68	0.00	4	0.58	0.00	6	0.51	0.43	9	0.79	0.10
DL032	7	0.73	0.27	2	0.50	1.00	2	0.40	0.00	1	0.00	0.00	4	0.50	0.40	5	0.59	0.36
SSR016	1	0.00	0.00	1	0.00	0.00	1	0.00	0.00	1	0.00	0.00	1	0.00	0.00	1	0.00	0.00
SSR014	13	0.79	0.26	1	0.00	0.00	4	0.58	0.11	5	0.51	0.40	6	0.64	0.43	9	0.77	0.19
SSR015	4	0.32	0.22	2	0.50	1.00	1	0.00	0.00	1	0.00	0.00	1	0.00	0.00	2	0.27	0.32
SSR017	9	0.72	0.09	1	0.00	0.00	1	0.00	0.00	2	0.16	0.20	6	0.68	0.14	7	0.66	0.08
SSR001	2	0.15	0.00	1	0.00	0.00	2	0.18	0.00	1	0.00	0.00	0	NA	NA	2	0.08	0.00
SSR003	3	0.25	0.00	1	0.00	0.00	2	0.41	0.00	2	0.28	0.00	1	0.00	0.00	2	0.25	0.00
SSR004	3	0.11	0.02	2	0.10	0.11	1	0.00	0.00	1	0.00	0.00	1	0.00	0.00	2	0.03	0.03
257	13	0.65	0.31	3	0.56	1.00	2	0.41	0.00	1	0.00	0.00	3	0.42	0.25	8	0.61	0.42
305	7	0.57	0.40	2	0.50	1.00	3	0.42	0.60	2	0.29	0.40	1	0.00	0.00	4	0.34	0.33
327	13	0.82	0.33	3	0.55	1.00	6	0.68	0.29	4	0.59	0.50	1	0.00	0.00	8	0.52	0.14
329	13	0.85	0.41	3	0.60	1.00	5	0.59	0.25	3	0.52	0.25	4	0.48	0.29	6	0.51	0.30
334	4	0.58	0.10	2	0.10	0.11	4	0.47	0.22	2	0.28	0.00	2	0.37	0.00	4	0.59	0.58
341	6	0.68	0.11	1	0.00	0.00	3	0.47	0.00	2	0.25	0.00	2	0.19	0.25	11	0.82	0.50
350	10	0.81	0.43	4	0.69	0.78	3	0.54	0.29	5	0.63	0.60	3	0.51	0.00	10	0.83	0.52
351	13	0.83	0.54	2	0.50	1.00	6	0.59	0.71	5	0.67	0.75	3	0.38	0.20	4	0.38	0.10
355	16	0.88	0.51	2	0.50	1.00	7	0.72	0.44	4	0.56	0.60	5	0.66	0.71	6	0.73	0.11
356	14	0.81	0.59	4	0.53	0.43	5	0.69	0.78	4	0.53	0.67	1	0.00	0.00	8	0.78	0.46
358	8	0.71	0.14	1	0.00	0.00	4	0.59	0.22	1	0.00	0.00	1	0.00	0.00	9	0.78	0.73
SSR009	10	0.66	0.39	1	0.00	0.00	4	0.62	0.56	3	0.50	0.50	6	0.69	0.75	13	0.87	0.71
SSR010	4	0.30	0.29	1	0.00	0.00	4	0.45	0.50	0	NA	NA	0	NA	NA	8	0.77	0.62
360	12	0.83	0.26	0	NA	NA	6	0.69	0.22	2	0.30	0.00	6	0.69	0.60	6	0.66	0.09
364	7	0.48	0.24	1	0.00	0.00	6	0.69	0.56	2	0.38	0.00	3	0.48	0.40	14	0.87	0.33
367	11	0.84	0.49	2	0.50	1.00	6	0.72	0.44	4	0.56	0.25	4	0.56	0.33	7	0.61	0.25
368	21	0.88	0.29	1	0.00	0.00	10	0.80	0.33	5	0.61	0.25	4	0.52	0.25	10	0.81	0.59
371	11	0.78	0.54	2	0.50	1.00	5	0.55	0.38	5	0.70	0.80	5	0.59	0.43	13	0.82	0.20
384	9	0.83	0.21	2	0.10	0.11	4	0.60	0.11	2	0.16	0.20	5	0.56	0.43	10	0.76	0.67
388	18	0.87	0.31	3	0.44	0.00	8	0.78	0.78	2	0.16	0.20	3	0.52	1.00	8	0.82	0.23
392	15	0.83	0.36	1	0.00	0.00	5	0.62	0.13	4	0.58	0.60	8	0.80	0.86	11	0.82	0.38
394	14	0.74	0.32	1	0.00	0.00	5	0.43	0.44	4	0.56	0.40	8	0.77	0.67	13	0.80	0.37
395	16	0.87	0.29	3	0.21	0.13	9	0.78	0.57	2	0.26	0.00	4	0.48	0.20	10	0.64	0.37
429	20	0.86	0.27	1	0.00	0.00	8	0.79	0.56	3	0.47	0.00	5	0.61	0.20	15	0.87	0.27
442	7	0.59	0.18	1	0.00	0.00	6	0.69	0.29	2	0.23	0.33	0	NA	NA	13	0.81	0.22
445	9	0.79	0.42	2	0.50	1.00	3	0.49	0.33	3	0.36	0.50	2	0.37	0.00	8	0.66	0.17

Table 3 (*concluded*).

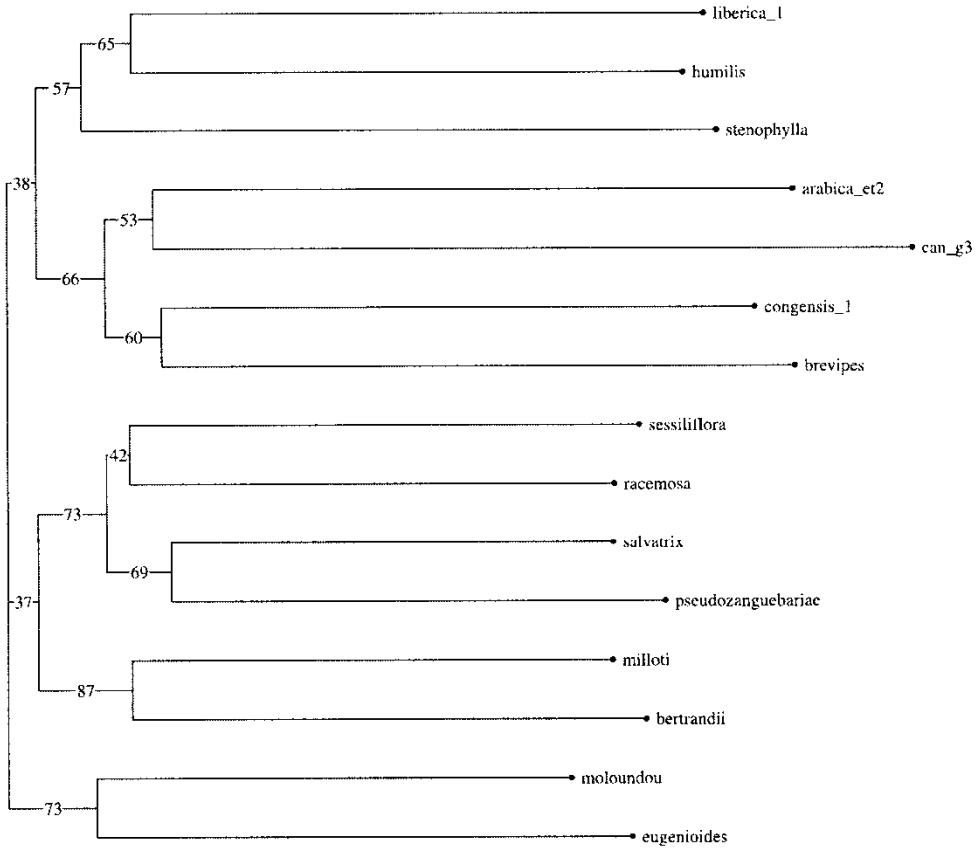
	15 species			*C. arabica*			*C. canephora*			*C. congensis*			*C. liberica*			4 species		
Marker	*N*	GD	H_o	*N*	GD	H_o	*N*	GD	H_o	*N*	GD	H_o	*N*	GD	H_o	*N*	GD	H_o
456	11	0.70	0.22	1	0.00	0.00	11	0.81	0.44	0	NA	NA	0	NA	NA	5	0.74	0.44
460	22	0.88	0.54	2	0.50	1.00	5	0.60	0.13	8	0.74	0.80	7	0.73	0.80	11	0.74	0.28
461	13	0.86	0.41	4	0.47	0.56	7	0.74	0.33	3	0.38	0.20	5	0.64	0.57	18	0.85	0.68
463	7	0.73	0.59	2	0.50	1.00	5	0.71	0.67	2	0.19	0.25	3	0.50	0.40	14	0.89	0.45
471	11	0.81	0.26	1	0.00	0.00	5	0.65	0.29	4	0.56	0.25	5	0.62	0.67	5	0.65	0.65
472	15	0.88	0.45	6	0.69	1.00	6	0.72	0.25	4	0.52	0.25	3	0.38	0.33	9	0.77	0.32
477	16	0.87	0.38	2	0.50	1.00	5	0.53	0.33	2	0.26	0.00	3	0.49	0.00	10	0.88	0.54
495	9	0.75	0.07	1	0.00	0.00	6	0.69	0.33	1	0.00	0.00	1	0.00	0.00	11	0.82	0.43
SSR005	11	0.69	0.10	2	0.10	0.11	3	0.50	0.00	3	0.29	0.20	3	0.45	0.20	7	0.66	0.10
501	16	0.85	0.47	2	0.49	0.89	9	0.79	0.56	1	0.00	0.00	5	0.51	0.57	14	0.87	0.58
753	13	0.82	0.58	3	0.54	1.00	5	0.66	0.38	4	0.65	1.00	4	0.53	0.67	8	0.79	0.72
755	15	0.87	0.59	3	0.59	1.00	9	0.76	0.56	5	0.67	0.75	4	0.60	0.80	13	0.89	0.76
774	8	0.58	0.12	2	0.10	0.11	3	0.26	0.11	1	0.00	0.00	1	0.00	0.00	6	0.53	0.10
779	9	0.86	0.58	2	0.50	1.00	7	0.74	0.38	4	0.58	0.60	5	0.73	0.86	9	0.85	0.71
782	9	0.77	0.19	5	0.68	0.80	1	0.00	0.00	5	0.64	0.20	4	0.54	0.20	6	0.73	0.25
790	16	0.88	0.56	3	0.60	1.00	9	0.77	0.67	5	0.66	0.60	4	0.42	0.43	14	0.86	0.71
809	8	0.71	0.53	2	0.50	1.00	3	0.35	0.44	1	0.00	0.00	5	0.72	1.00	7	0.67	0.65
837	10	0.82	0.24	3	0.29	0.13	6	0.70	0.43	3	0.42	0.25	3	0.48	0.20	9	0.82	0.25
838	16	0.89	0.46	3	0.60	1.00	5	0.65	0.22	2	0.28	0.50	3	0.45	0.20	10	0.87	0.50
Mean*	11	0.72	0.32	2	0.30	0.49	5	0.55	0.29	3	0.34	0.27	4	0.44	0.34	8	0.69	0.37
Mean†	11	0.72	0.32	2	0.30	0.48	5	0.55	0.30	3	0.35	0.27	4	0.45	0.35	8	0.69	0.37
SD	1	0.03	0.02	0	0.03	0.06	0	0.03	0.03	0	0.03	0.04	0	0.03	0.04	0	0.03	0.03
2.5% l.b.	10	0.66	0.27	2	0.23	0.37	4	0.49	0.24	2	0.29	0.20	3	0.39	0.27	7	0.63	0.31
97.5% u.b.	12	0.76	0.36	2	0.36	0.61	5	0.60	0.35	3	0.41	0.34	4	0.51	0.43	9	0.74	0.42

Note: *N*, number of alleles; GD, gene diversity; H_o, observed heterozygosity; NA, not available (missing data); SD, standard deviation; 2.5% l.b. and 97.5% u.b., lower and upper boundaries of the 95% confidence interval.

*Mean values based only on markers with no missing data for the considered species.

†Mean values calculated over 5000 bootstrap iterations and based only on markers with no missing data for the considered species.

Fig. 2. Neighbor-joining tree for 15 individuals (one per species) based on the dissimilarity matrix calculated by simple matching. Bootstrap values were calculated with 5000 repetitions.



C. arabica sequences, since mean percentages of amplification are the same. This result will be used for development of comparative mapping, utilization of new markers, and knowledge transfer from one species to another.

SSRs described in genes involved in sucrose metabolism appear to have some specific behaviour, since they exhibit very low diversity (1–4 alleles in the global sample) or intermediate diversity (9–13 alleles). These results will allow us to use these markers to study gene regions implicated in sucrose metabolism.

In our work, using new markers, we validate the relation between *C. anthonyi* and *C. eugenioides*, which was previously described by Lashermes et al. (1997). These two species show high similarity based on both morphological and molecular data. However, *C. anthonyi* originated from Cameroon, while *C. eugenioides* is native to East Africa. No other coffee species belonging to the same clade (C) has been observed between these two distant geographic areas, and there is no clear explanation for the discontinuous distribution of these coffee trees (Anthony 1992).

We can use these two species to improve *C. arabica* varieties, considering their genetic relationships and the original self-compatible system of *C. anthonyi* (Anthony et al. 2006). These two species show some of the lowest concentrations of caffeine (0.6%) of the genus *Coffea* and exhibit high concentrations of trigonelline (1.6% for *C. anthonyi*, 1.3% for *C. eugenioides*; F. Anthony, personal communication), an alkaloid compound. These two characters have always interested breeders in coffee improvement. Meanwhile, since few genotypes are in collection worldwide, these two species have not been agronomically well characterized and experiments are necessary to assess potential resistances to biotic and abiotic stresses usable for improvement.

On the other hand, part of the *C. arabica* genome has been shown to originate from an ancestral species genetically close to *C. eugenioides* or *C. anthonyi* (Lashermes et al. 1999). These relationships can be used to better understand the elaboration and functioning of the allotetraploid genome of *C. arabica*, in particular comportment of homeologous chromosomes during meiosis.

Diversity and genetic properties of cultivated and related wild species

The diversity and genetic relationships of *C. arabica*, *C. canephora*, and related species are examined in our work. *Coffea arabica* has been treated as a diploid species because of the presence of only 2 alleles on all the loci. This is not surprising considering the allotetraploid origin and amphidiploid nature of *C. arabica* and its autogamy. *Coffea arabica* is the only species that exhibited an expected heterozygosity lower than the observed heterozygosity. This result is consistent with other studies (Lashermes et al. 1999; Aggarwal et al. 2007). It could result from the fixed heterozygosity (Lashermes et al. 1999) during the speciation process including two different ancestral genomes. Data derived from SNP analysis (Pot et al. 2006) confirm this hypothesis with the construction of two haplotypes based on sequences. One is close to *C. canephora* and related species,

while the other exhibits strong relationships with *C. eugenioides*. However, heterozygosity within the two ancestral genomes appears to have been lost, since only one allele from each genome remains in *C. arabica*. This result indicates a possible lack of recombination between the ancestral genomes, while recombination within each genome occurs normally.

We included the two varieties of *C. liberica*, i.e., *C. liberica* var. *liberica* and *C. liberica* var. *dewevrei*. These two varieties were genetically well differentiated in previous work (N'Diaye et al. 2005). In our study, the differentiation between these two varieties and their divergence from other species was confirmed.

Coffea congensis, which is considered an ecotype of *C. canephora* (Prakash et al. 2005), is differentiated from *C. canephora*, but both species are grouped in the same cluster in Fig. 2. Our study also points out the relatedness of *C. canephora* and *C. brevipes*. *Coffea brevipes* originated from Cameroon and Gabon (Chevalier 1947; Anthony 1992; Stoffelen 1998). This species has been described, like *C. congensis* (Sybenga 1960; Anthony 1992; Prakash et al. 2005), as an ecotype of *C. canephora* (Chevalier 1947; Anthony 1992; Stoffelen 1998). Our work provides evidence to confirm the hypothesis that *C. brevipes* is a dwarf form of *C. canephora*, since this species appears to be related to the Central African genotypes of *C. canephora* (Fig. 1). Field studies should be performed to validate this point of view.

Coffea canephora is the most diverse species, with 95 private alleles, i.e., 31.25% of the total number of private alleles and 14.66% of the total number of alleles. Our results (Fig. 1) confirm the division of this species into at least two groups, i.e., a Congolese group from Central Africa and a Guinean group from West Africa. In contrast, *C. liberica* and *C. congensis* exhibit, respectively, 52 and 27 private alleles, while *C. arabica* presents 20 private alleles. The global amount of diversity for *C. canephora*, *C. congensis*, and *C. liberica* is very high compared with that for *C. arabica*, which has the lowest diversity even if wild individuals of this species are more diverse than cultivated ones. These results are in accordance with previous studies (Anthony et al. 2002*a*; Moncada and McCouch 2004) and corroborate the very narrow genetic base of *C. arabica*, suggesting a small number of founders for this species.

Conclusion and consequences for breeding

Our work shows the transferability of SSR markers over the genus *Coffea*. We point out the potential usefulness of related wild species in breeding strategies for *C. arabica* and *C. canephora* to provide new variability. These results increase the importance of genus diversity studies. Our results, as well as previous analyses using ITS and RFLP markers (Lashermes et al. 1997, 1999), lead us to consider that a high potentiality for breeding has not yet been exploited using species of these two clades.

We propose working on two axes. First, since *C. liberica*, *C. congensis*, and the cultivated species are all grouped in related clades, the potentialities of crosses between these species are high and the resulting hybrids would have an important level of fertility (Louarn 1992). Variability observed within these species can be used for improvement of beverage and bean quality, productivity, and resistance to biotic and abiotic stresses in the cultivated species. Second, breeding potentialities with species from other diversity groups are important to assess, since interesting characters have been described. For example, *C. racemosa* (E clade according to Cros et al. 1998) has been used for coffee leaf miner resistance (Guerreiro et al. 1999; Mondego et al. 2005) and *C. anthonyi* (C clade) could be used for self-compatibility.

Breeding *C. arabica* will have to take into account its allopolyploid origin. Considering the low rate of recombination between the two ancestral genomes, the introduction of recessive alleles coding for traits of interest will be difficult.

Comparative genetic mapping and association mapping will be developed for future breeding programs. Relationships between *C. canephora*, *C. eugenioides*, *C. arabica*, and related species will be analysed to assess valuable traits for both quality and resistance improvement throughout the genus.

Acknowledgements

Technical help was provided by the Montpellier Languedoc-Roussillon Genopole genotyping platform. The authors thank the NARO-CORI (Uganda), the CNRA (République de Côte d'Ivoire), and the IRD (France) for providing plant material. P. Cubry is supported by a grant of the French ministry of research. The authors are grateful to J.L. Noyer for discussions and advice on an early version of the manuscript. We also thank an anonymous reviewer for comments and advice on this paper.

References

Aggarwal, R.K., Hendre, P.S., Varshney, R.K., Bhat, P.R., Krishnakumar, V., and Singh, L. 2007. Identification, characterization and utilization of EST-derived genic microsatellite markers for genome analyses of coffee and related species. Theor. Appl. Genet. **114**: 359–372. PMID:17115127.

Alvarez, A.E., van de Wiel, C.C.M., Smulders, M.J.M., and Vosman, B. 2001. Use of microsatellites to evaluate genetic diversity and species relationships in the genus *Lycopersicon*. Theor. Appl. Genet. **103**: 1283–1292. doi:10.1007/s001220100662.

Anthony, F. 1992. Les ressources génétiques des caféiers: collecte, gestion d'un conservatoire et évaluation de la diversité génétique. Collection Travaux et Documents Microfichés n° 81, ORSTOM (now IRD), Paris.

Anthony, F., Combes, C., Astorga, C., Bertrand, B., Graziosi, G., and Lashermes, P. 2002*a*. The origin of cultivated *Coffea arabica* L. varieties revealed by AFLP and SSR markers. Theor. Appl. Genet. **104**: 894–900. PMID:12582651.

Anthony, F., Quirós, O., Topart, P., Bertrand, B., and Lashermes, P. 2002*b*. Detection by simple sequence repeat markers of introgression from *Coffea canephora* in *Coffea arabica* cultivars. Plant Breed. **121**: 542–544. doi:10.1046/j.1439-0523.2002.00748.x.

Anthony, F., Noirot, M., Couturon, E., and Stoffelen, P. 2006. New coffee (*Coffea* L.) species from Cameroon bring original characters for breeding [CD-ROM]. *In* 21st International Conference on Coffee Science, Montpellier, 11–15 September 2006. *Edited by* ASIC. Paris, France.

Berthaud, J. 1986. Les ressources génétiques pour l'amélioration des caféiers africains diploïdes. Doctoral thesis, Université de Paris-Sud, Orsay, France.

Casasoli, M. 2004. Cartographie génétique comparée chez les faga-

cées. Doctoral thesis, Université de Bordeaux 1, Bordeaux, France.

Chevalier, A. 1947. Les caféiers du globe, fascicule III, Systématique des caféiers et faux caféiers. Paris.

Combes, M.C., Andrzejewski, S., Anthony, F., Bertrand, B., Rovelli, P., Graziosi, G., and Lashermes, P. 2000. Characterization of microsatellite loci in *Coffea arabica* and related coffee species. Mol. Ecol. **9**: 1178–1180. doi:10.1046/j.1365-294x.2000.00954-5.x. PMID:10964241.

Cramer, P.J.S. 1948. Les caféiers hybrides du groupe Congusta. Bull. Agric. du Congo Belge, **34**: 29–48.

Cros, J., Combes, M.C., Trouslot, P., Anthony, F., Hamon, S., Charrier, A., and Lashermes, P. 1998. Phylogenetic analysis of chloroplast DNA variation in *Coffea* L. Mol. Phylogenet. Evol. **9**: 109–117. doi:10.1006/mpev.1997.0453. PMID:9479700.

Cubry, P., De Bellis, F., Pot, D., Musoli, P., Legnaté, H., Leroy, T., and Dufour, M. 2005. Genetic diversity analyses and linkage disequilibrium evaluation in some natural and cultivated populations of *Coffea canephora*. *In* Proceedings of the 4th Plant Genomics European Meeting, Amsterdam, 20–23 September 2005.

Davis, A., and Stoffelen, P. 2006. An annotated taxonomic conspectus of the genus *Coffea* (Rubiaceae). Bot. J. Linn. Soc. **152**: 465–512. doi:10.1111/j.1095-8339.2006.00584.x.

Dufour, M., Hamon, P., Noirot, M., Risterucci, A.M., and Leroy, T. 2001. Potential use of SSR markers for *Coffea* spp. genetic mapping [CD-ROM]. *In* 19th International Scientific Colloquium on Coffee, Trieste, 2001. *Edited by* ASIC. Paris, France.

Dussert, S., Lashermes, P., Anthony, F., Montagnon, C., Trouslot, P., Combes, M.-C., et al. 2003. Coffee (*Coffea canephora*). *In* Genetic diversity of cultivated tropical plants. *Edited by* P. Hamon, M. Seguin, X. Perrier, and J.C. Glaszmann. Science Publishers, Inc., Enfield, N.H. pp. 239–258.

Gao, L.Z., Zhang, C.H., and Jia, J.Z. 2005. Cross-species transferability of rice microsatellites in its wild relatives and the potential for conservation genetic studies. Genet. Resour. Crop Evol. **52**: 931–940. doi:10.1007/s10722-003-6124-3.

Geromel, C., Ferreira, L.P., Cavalari, A.A., Pereira, L.F.P., Guerreiro, S.M.C., Vieira, L.G.E., et al. 2006. Biochemical and genomic analysis of sucrose metabolism during coffee (*Coffea arabica*) fruit development. J. Exp. Bot. **57**: 3243–3258. doi:10.1093/jxb/erl084. PMID:16926239.

Guerreiro, O., Silvarolla, M.B., and Eskes, A.B. 1999. Expression and mode of inheritance of resistance in coffee to leaf miner *Perileucoptera coffeella*. Euphytica, **105**: 7–15.

The International Plant Names Index. 2007. Available from http://www.ipni.org [accessed 10 December 2007].

Jarne, P., and Lagoda, P.J. 1996. Microsatellites, from molecules to populations and back. Trends Ecol. Evol. **11**: 424–429. doi:10.1016/0169-5347(96)10049-5.

Lashermes, P., Cros, J., Combes, M.C., Trouslot, P., Anthony, F., Hamon, S., and Charrier, A. 1996. Inheritance and restriction fragment length polymorphism of chloroplast DNA in the genus *Coffea* L. Theor. Appl. Genet. **93**: 626–632.

Lashermes, P., Combes, M.C., Trouslot, P., and Charrier, A. 1997. Phylogenetic relationships of coffee-tree species (*Coffea* L.) as inferred from ITS sequences of nuclear ribosomal DNA. Theor. Appl. Genet. **94**: 947–953. doi:10.1007/s001220050500.

Lashermes, P., Combes, M.C., Robert, J., Trouslot, P., D'Hont, A., Anthony, F., and Charrier, A. 1999. Molecular characterization and origin of the *Coffea arabica* L. genome. Mol. Gen. Genet. **261**: 259–266. PMID:10102360.

Leroy, T., Marraccini, P., Dufour, M., Montagnon, C., Lashermes, P., Sabau, X., et al. 2005. Construction and characterization of a *Coffea canephora* BAC library to study the organization of sucrose biosynthesis genes. Theor. Appl. Genet. **111**: 1032–1041. doi:10.1007/s00122-005-0018-z. PMID:16133319.

Liu, K., and Muse, S.V. 2005. PowerMarker: integrated analysis environment for genetic marker data. Bioinformatics, **21**: 2128–2129. doi:10.1093/bioinformatics/bti282. PMID:15705655.

Louarn, J. 1992. La fertilité des hybrides interspécifiques et les relations génomiques entre caféiers diploïdes d'origine africaine (genre *Coffea* L., sous-genre *Coffea*). Doctoral thesis, Université de Paris-Sud, Orsay, France.

Moncada, P., and McCouch, S. 2004. Simple sequence repeat diversity in diploid and tetraploid *Coffea* species. Genome, **47**: 501–509. doi:10.1139/g03-129. PMID:15190367.

Mondego, J.M.C., Guerreiro-Filho, O., Bengtson, M.H., Drummond, R.D., Felix, J.M., Duarte, M.P., et al. 2005. Isolation and characterization of *Coffea* genes induced during coffee leaf miner (*Leucoptera coffeella*) infestation. Plant Sci. **69**: 351–360.

Montagnon, C. 2000. Optimization des gains génétiques dans le schéma de séléction récurrente réciproque de *Coffea canephora* Pierre. Doctoral thesis, Ecole Nationale Supérieure Agronomique de Montpellier, Montpellier, France.

Musoli, P., Aluka, P., Cubry, P., Dufour, M., De Bellis, F., Ogwang, J., et al. 2006. Fighting against coffee wilt disease: Uganda wild canephora genetic diversity and usefulness. *In* 21st International Conference on Coffee Science, Montpellier, 11–15 September 2006. *Edited by* ASIC. Paris, France.

N'Diaye, A., Poncet, V., Louarn, J., Hamon, S., and Noirot, M. 2005. Genetic differentiation between *Coffea liberica* var. *liberica* and *C. liberica* var. *dewevrei* and comparison with *C. canephora*. Plant Syst. Evol. **253**: 95–104. doi:10.1007/s00606-005-0300-1.

Perrier, X., Flori, A., and Bonnot, F. 2003. Data analysis methods. *In* Genetic diversity of cultivated tropical plants. Science Publishers, Inc., Enfield, N.H. pp. 43–76.

Poncet, V., Hamon, P., Minier, J., Carasco, C., Hamon, S., and Noirot, M. 2004. SSR cross-amplification and variation within coffee trees (*Coffea* spp.). Genome, **47**: 1071–1081. doi:10.1139/g04-064. PMID:15644965.

Poncet, V., Dufour, M., Hamon, P., Hamon, S., de Kochko, A., and Leroy, T. 2007. Development of genomic microsatellite markers in *Coffea canephora* and their transferability to other coffee species. Genome, **50**(12):1156–1161. doi:10.1139/G07-073.

Pot, D., Bouchet, S., Cubry, P., Dufour, M., De Bellis, F., Jourdan, I., et al. 2006. Nucleotide diversity of genes involved in sucrose metabolism. Towards the identification of candidate genes controling sucrose variability in *Coffea* spp. *In* 21st International Conference on Coffee Science, Montpellier, 11–15 September 2006. *Edited by* ASIC. Paris, France.

Prakash, N.S., Combes, M.C., Somanna, N., and Lashermes, P. 2002. AFLP analysis of introgression in coffee cultivars (*Coffea arabica* L.) derived from a natural interspecific hybrid. Euphytica, **124**: 265–271. doi:10.1023/A:1015736220358.

Prakash, N.S., Combes, M.C., Dussert, S., Naveen, S., and Lashermes, P. 2005. Analysis of genetic diversity in Indian robusta coffee genepool (*Coffea canephora*) in comparison with a representative core collection using SSRs and AFLPs. Genet. Resour. Crop Evol. **52**: 333–343. doi:10.1007/s10722-003-2125-5.

Risterucci, A.M., Grivet, L., N'Goran, J.A.K., Pieretti, I., Flament, M.H., and Lanaud, C. 2000. A high-density linkage map of *Theobroma cacao* L. Theor. Appl. Genet. **101**: 948–955. doi:10.1007/s001220051566.

Rovelli, P., Mettulio, R., Anthony, F., Anzueto, F., and Lashermes, P. 2000. Microsatellites in *Coffea arabica* L. *In* Coffee biotech-

nology and quality. *Edited by* T. Sera, C.R. Soccol, A. Pandey, and S. Roussos. Kluwer Academic Publishers, the Netherlands. pp. 123–133.

Rozen, S., and Skaletski, H.J. 2000. Primer 3. Version 0.2 [computer program]. Available from http://primer3.sourceforge.net/.

Saitou, N., and Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**: 406–425. PMID:3447015.

Stoffelen, P. 1998. *Coffea* and *Psilanthus* (Rubiaceae) in tropical Africa: a systematic and palynological study, including a revision of the West and Central African species. Doctoral thesis, Katholieke Universiteit Leuven, Leuven, Belgium.

Sybenga, J. 1960. Genetics and cytology of coffee. A literature review. Bibliographica Genet. **19**: 217–316.

Tautz, D., and Renz, M. 1984. Simple sequences are ubiquitous repetitive components of eukaryotic genomes. Nucleic Acids Res. **12**: 4127–4138. doi:10.1093/nar/12.10.4127. PMID:6328411.

Yu, K., Park, S.J., and Poysa, V. 1999. Abundance and variation of microsatellite DNA sequences in beans (*Phaseolus* and *Vigna*). Genome, **42**: 27–34. doi:10.1139/gen-42-1-27.

Supplementary data : voir Annexes A.2.1 à A.2.4

Conclusion sur la diversité du genre *Coffea*

Ce travail nous a permis de confirmer et préciser la place qu'occupe *C. canephora* au sein du genre *Coffea*. Nous avons ainsi pu confirmer sa place de choix dans un clade à proximité de nombreuses autres espèces déjà étudiées ou utilisées en amélioration telle que *C. congensis* ou *C. liberica*. Ces proximités phylogénétiques, déjà envisagées au vu du nombre important d'hybrides naturels possibles ou d'hybrides artificiels facilement obtenables, nous permettent d'envisager la possibilité de généraliser nos futures études menées sur *C. canephora* à l'ensemble du clade WC (caféiers du Centre et de l'Ouest de l'Afrique).

Cette proximité génétique entre espèces peut s'expliquer par la jeunesse du genre *Coffea*. De plus, ces résultats sont concordants avec les études réalisées sur la fertilité des hybrides interspécifiques, suggérant l'existence d'un complexe d'espèces plutôt que d'espèces réellement isolées au sein du genre, avec des flux de pollen occasionnels entre certaines espèces (Louarn, 1992).

La proximité génétique de *C. canephora* avec une partie du génome de *C. arabica* facilite les processus d'introgression et des hybrides naturels fertiles peuvent exister. L'hybride de Timor en est un exemple particulièrement intéressant, avec une assez bonne qualité à la tasse et des résistances importantes à des stress biotiques et abiotiques, d'autres hybrides naturels peuvent être observés, notamment en Nouvelle-Calédonie, où les 2 espèces ont été introduites. Les résultats de nos travaux permettent d'envisager de manière intéressante des possibilités de suivi de ces introgressions de régions génomiques et de caractères d'intérêt de *C. canephora* vers *C. arabica*. Les approches d'amélioration de *C. arabica* par *C. canephora* déjà existantes, notamment par la création d'hybrides arabusta, pourront être affinées et améliorées par l'utilisation de marqueurs moléculaires. Des études plus poussées du génome de *C. arabica* seront nécessaires et le recours à ces marqueurs moléculaires, spécialement les microsatellites, permettra de suivre de manière précise la manière dont se déroulent les recombinaisons, en testant notre capacité à induire une introgression non seulement sur la partie du génome de *C. arabica* proche de *C. canephora* mais également sur celle proche de *C. eugenioides*. *C. arabica* étant une plante allotétraploïde montrant une méiose de type amphidiploïde (Lashermes *et al.*, 1999), ces questions sont particulièrement importantes à résoudre pour envisager de manière efficace les possibilités d'extension de nos recherches sur *C. canephora* à *C. arabica*.

Un résultat plus immédiat de notre étude pourrait être l'identification de zones génomiques d'intérêt voire de polymorphismes causaux de variations de caractères d'intérêt chez *C. canephora*, qui pourrait être élargie à *C. arabica*. L'intérêt principal de telles approches est la plus grande facilité de travail et de marquage moléculaire sur une espèce diploïde que sur une espèce polyploïde.

Cette étude a aussi permis d'évaluer de manière comparative la diversité de *C. canephora* vis-à-vis des autres espèces du genre, représentées par un échantillon représentatif de leur diversité connue. *C. canephora* apparaît comme l'espèce la plus diverse, avec une très forte structure génétique. La place taxonomique des caféiers de la Nana (groupe C) a longtemps fait l'objet d'interrogations pour les sélectionneurs et taxonomistes (Berthaud & Guillaumet, 1978), nous confirmons dans ce travail la place de cette population au sein de l'espèce *C. canephora*.

Nous avons donc clairement précisé la place de notre espèce au sein du genre, montré sa très importante diversité et sa forte structure génétique, et posé les bases de notre capacité à généraliser les résultats qui seront obtenus dans la suite de notre travail.

Chapitre 3 : La diversité intra-spécifique de *Coffea canephora* étudiée à partir d'un panel de génotypes issus de collections

Introduction

Coffea canephora est une espèce qui possède une aire de répartition très importante, couvrant une grande partie des régions intertropicales d'Afrique. Sa culture est de plus étendue à l'ensemble des régions intertropicales mondiales (Afrique, Amérique et Asie). Nous avons confirmé dans le précédent chapitre que la diversité et l'hétérozygotie de cette espèce sont globalement élevées, ce qui avait été avancé par Berthaud (1986) et Dussert *et al.* (1999). De plus, Berthaud (1986) a décrit une structuration importante de cette diversité, structuration ensuite précisée (Montagnon *et al.*, 1992 ; Dussert *et al.*, 1999), et qui apparaît également dans notre étude du chapitre 2. Cette structure s'accompagne, comme nous l'avons déjà évoqué dans l'introduction, d'une importante variabilité phénotypique et d'adaptations importantes, allant par exemple d'arbres très sensibles à la sécheresse à des arbres possédant des adaptations physiologiques telles que la fermeture des stomates ou le repli des feuilles permettant une tolérance importante à des épisodes de sécheresse (Montagnon & Leroy, 1993).

D'un point de vue général, la connaissance de la diversité et de la structure génétique est un enjeu primordial sur les espèces que l'on souhaite étudier, soit à des fins d'amélioration à travers la connaissance et la caractérisation des ressources génétiques potentiellement disponibles, soit à des fins de conservation pour certaines espèces classifiées comme menacées.

La structure génétique, façonnée par de multiples évènements démographiques et processus évolutifs, est particulièrement importante à connaître si l'on veut mener des études de déséquilibre de liaison et par extension des études de génétique d'association (Pritchard *et al.*, 2000a; Flint-Garcia *et al.*, 2005). La recherche d'associations entre des marqueurs et des caractères d'intérêt agronomique pour l'amélioration est en effet l'un des principaux outils à

disposition pour optimiser les processus de sélection des plantes, en assistant l'amélioration grâce à des marqueurs moléculaires liés à des traits d'intérêt, rendant plus efficace et plus rapide le progrès génétique. Cette recherche peut se faire principalement selon deux stratégies, les études d'association comme citées précédemment et la cartographie génétique complétée par la recherche de QTL.

Le but de ce chapitre est donc dans un premier temps d'étudier un large panel d'individus à l'aide de marqueurs microsatellites répartis sur l'ensemble du génome, afin de pouvoir étudier au mieux la diversité et la structure génétique de notre espèce à un niveau global. Nous avons caractérisé un total de 293 individus à l'aide de 39 microsatellites et nous avons soumis nos données à une analyse de diversité, suivie d'une analyse poussée de la structure génétique faisant appel aux deux grandes méthodes d'études de celle-ci (les modèles de génétique des populations et les analyses multivariées). Ces analyses confirment et précisent les résultats précédemment rapportés par Berthaud et Dussert et dressent les bases d'une future caractérisation des collections.

Dans une seconde partie, nous proposerons et validerons un panel de marqueurs pour analyser des génotypes inconnus et les situer dans la diversité de l'espèce, validant notre approche dans le but de caractériser rapidement les collections à l'aide d'un nombre raisonnable de marqueurs, dans une stratégie proche de celle qui a été développée par exemple chez la vigne *Vitis vinifera* (Le Cunff *et al.*, 2008).

Une troisième partie traitera de l'importance de la cartographie génétique et des résultats obtenus par la recherche de QTL. Les régions génomiques ainsi identifiées pourront servir de cible à de futures études d'association ou de sélection nucléotidique.

Nous discuterons enfin de l'implication de ces études sur les possibilités de réaliser une ou des core-collections de *C. canephora*, de l'identification de populations ou groupes de diversité cibles pour des études préliminaires de déséquilibre de liaison et nous relierons cette dernière proposition avec des résultats de cartographie génétique.

Manuscrit soumis à Annals of Botany

Diversité et structure des populations de *Coffea canephora* (Rubiaceae) analysées par microsatellites.

Original Article

Diversity and population structure of *Coffea canephora* (Rubiaceae) assessed by microsatellites.

Philippe Cubry[*1], Fabien De Bellis[1], Pascal Musoli[2], David Pot[1] and Thierry Leroy[1]

[1] CIRAD, UMR DAP, TA A-96/03, Avenue Agropolis, Montpellier, F-34398 France
[2] Coffee Research Institute, P.O. Box 185, Mukono, Uganda

Keywords: genetic diversity, population genetics, differentiation, structure, refuge zone, tropical areas, *Coffea canephora*, Ivory Coast, Uganda.

[*]Corresponding author: Philippe Cubry, TA A-96/03, Avenue Agropolis, 34398 Montpellier Cedex 5, France.
Telephone: (33) 4-67-61-56-90
Fax: (33) 4-67-61-57-93
Email: philippecubry@orange.fr

Running title: genetic diversity and population structure of *Coffea canephora*

Keywords: *Coffea canephora*, robusta coffee, perennials, microsatellites, population structure, diversity, linkage disequilibrium, collection, genetic resources, diversity

Abstract

Background and aims This study examined the pattern of genetic variability and genetic relationships of cultivated, wild and improved populations of *Coffea canephora* Pierre ex. Frohener sampled throughout the species' area of distribution. The aim of this study was to confirm and fine-tune the previous structure of this species using new markers, and to identify a target population to undertake Linkage Disequilibrium studies.

Methods A total of 293 individuals was genotyped with 39 nuclear microsatellite markers. Genetic diversity and structure was investigated with both a model-based and a graphic approach. Genetic and geographical relationships were analysed and a phylogeography study was launched.

Key results An overall structure of 6 clusters was found. On the finest level of analysis, the model-based approach appeared to perform poorly compared to graphical and statistical analysis. Clear isolation by distance was found on a continent level but not on a country level in Ivory Coast. Relationships between the identified diversity clusters are discussed with regard to differentiation due to several glacial refuges during the Last Glacial Maximum. Several populations suitable for LD analyses were identified.

Conclusions We confirm and fine-tuned the genetic structure of the *Coffea canephora* species for a large sample of individuals and markers. Use of genotypes from collections outside Africa revealed the relatively narrow bases of diversity usable for breeding outside Ivory Coast. Future studies, including association mapping, should use results from this genetic structure study to limit the number of genotypes to be phenotypes in order to reduce the cost of collection management. We also give a basis for future characterisation of the whole reference collection of *C. canephora*. This study also provides the potential for determining the origin of genotypes in collections worldwide.

Introduction

Knowledge on genetic diversity in crops is a competitive task for both conservation and breeding purposes. A better understanding of genetic diversity is of a great importance for breeding requirements, such as tolerance of drought stress related to climatic changes (Tuberosa and Salvi, 2006) or quality improvement. Association studies in perennials (Neale, 2007) are also based on such diversity studies and can help to reduce the time needed to obtain new improved varieties. However, the number of false positives induced by genetic structure reinforces the needs to know that structure and take it into account in such studies.

The genus *Coffea*, from the Rubiaceae family, consists of 103 species (Davis and Stoffelen, 2006), originating from the intertropical regions of Africa and Madagascar. Within that diversity, only two species are commonly grown to produce commercial coffee, *Coffea arabica* L and *Coffea canephora* Pierre.

Robusta coffee produced by *Coffea canephora* Pierre (ex. Froehner) accounts for about 35 to 40% of total coffee production (International Coffee Organization, www.ico.org). *C. canephora* is cultivated at low to medium altitudes in the intertropical regions of Africa, America and Indonesia. *C. canephora* is a self-incompatible, diploid species indigenous to some areas of the tropical African forest, stretching from West Africa through Cameroon, Central African Republic (CAR), Congo, the Democratic Republic of Congo (DRC), Uganda, and northern Tanzania up to northern Angola. *C. canephora* populations are generally small disconnected ones with a small number of mother trees and few offspring scattered over small areas (less than 1 ha). Pollen dispersal may occur over a few metres up to a few kilometres, with competition between local and allogenous pollen, implying that pollination occurs mainly between trees belonging to the same population rather than between trees from different populations. Seed dispersal might occur over larger distances (up to a hundred kilometres) through the action of mammals or birds (Berthaud, 1986).

Previous diversity analyses based on isozyme (Berthaud, 1986) and RFLP (Dussert *et al.*, 1999) markers highlights the great genetic diversity of *C. canephora*. Phenotypic studies give evidence for wide genetic and phenotypic diversity structured into at least 2 main groups (Montagnon *et al*., 1992; Montagnon, 2000). Those main groups include a Guinean (G) and a Congolese group. The Congolese group can be subdivided into four smaller ones (i.e. SG2 and B from the Congo basin, C from the Central Africa Republic and Cameroon, and SG1 from the Atlantic coastal region of central Africa). SG1 and G are mainly small trees tolerant of drought, while C, B and SG2 are bigger trees with large leaves and are more susceptible to drought. A recent study added a new diversity group to the Congolese group, with wild individuals from Uganda (Musoli *et al.*, unpublished).

Worldwide collections of *C. canephora* contain a large number of genotypes derived from different surveys throughout the area of distribution and need to be characterized to assess the redundancy and capabilities of each collection. In our study, we analysed the genetic diversity of *C. canephora* using microsatellite markers on a large set of genotypes, most of which were already classified on the five described groups. We also analysed some new genotypes from various field collections in an attempt to find their genetic origin. Since *C. canephora* has only been grown since the late 19th century or the early 1900s and has a generation time of

about ten to 30 years, there is no clear domestication syndrome. The populations that have been studied, even cultivated, are close or identical to the wild original populations from which they were imported, with perhaps a few exceptions for some improved material (Montagnon, 2000).

This study was undertaken to confirm and fine-tune the previously investigated structure using new markers. We tried to understand the relationships between the diversity groups and the way the genetic structure was built. Finally, we identified several natural and composite populations to perform Linkage Disequilibrium (LD) studies in unstructured samples.

Materials and methods

Plant material and sampling

A total of 293 individuals from *Coffea canephora* was genotyped in this study. The aim of our sampling work was to cover the whole natural range of this species as defined by (Berthaud, 1986) and represent the different diversity groups already known (Table 3.1 and Figure 3.1, a complete list of genotypes is provided as supplementary material). Wild, cultivated and improved varieties were sampled. Several wild populations and a cultivated population from Ivory Coast (IC) were analysed, while other origins were represented by only one or two populations.

The wild Guinean populations were collected by ORSTOM (Office de la Recherche Scientifique et Technique d'Outre-Mer) and IRCC (Institut de Recherche du Café et du Cacao) in the primary and secondary forests of IC and Guinea from 1975 to the 1990s (Berthaud, 1986; Montagnon *et al*., 1992).

The cultivated genotypes from IC were collected from farms where hybridisation may have occurred between wild native Guinean and imported Congolese varieties from Central Africa (Montagnon *et al.*, 1993).

The Niaouli accessions were grown from seeds of cultivated trees originating from Benin (Montagnon, CIRAD, Montpellier, France, pers. comm.). However, it is likely that this variety came from the Noyo region of the Gabon Atlantic coast or from the Kouilou river region of Congo (Adibolo and Bertrand, 1988; Bodard, 1965, Portères, 1937).

The Nana population was surveyed in the Central African Republic (CAR) in Nana-Mambéré province, by an ORSTOM mission in 1975, near Dongué (Berthaud and Guillaumet, 1978; Berthaud et al., 1984).

The Libengé population was collected in primary forest on the island of Libengé, on the border between CAR and the Democratic Republic of Congo (DRC) (Berthaud and Guillaumet, 1978).

The Ugandan cultivated accessions (erect and nganda forms) were likely to be improved varieties, disseminated by INEAC (Institut National d'Etudes Agronomiques du Congo belge) from its Yangambi centre (Bodard, 1965; Thomas, 1935; Musoli *et al*., unpublished) through Java. INEAC 2 and 7 were also derived from that origin. These genotypes were offspring from two mother-trees planted at the Duekoue station (Ivory Coast) and imported from INEAC's Yangambi centre in 1935 (Portères, 1959; Bodard, 1965). Origin data for these genotypes were based exclusively on manuscript notes.

The Nemaya trees were genotypes with resistance to nematodes, used to produce resistant hybrids, and were obtained from the CATIE collection. Data on the collection indicated that one of these genotypes was imported from DRC (sp43), while the other was imported from Indonesia (sp44). Morphologically, those trees seemed to be related to the Congolese Group (Charmetant, CIRAD, Montpellier, France, pers. comm.).

Genotypes from Brazilian collections, as well as some unknown genotypes, were not well documented. Guyana accessions comprised 2 genotypes (Guy1 and Guy2), which were hybrids between two previously known genetic groups, the C and G groups. Guy3 came from a plantation survey, as well as the cultivated Ivorians, while Guy4 was an improved genotype.

All these genotypes, except the Ugandan, Brazilian and Guyana origins were obtained from the Divo reference collection by courtesy of CNRA under a CIRAD ATP (Programmed Thematic Project). Ugandan (erect and nganda) accessions, as well as wild individuals (uw), were obtained courtesy of Naro-CORI in Uganda and have been more closely characterized in other work (Musoli *et al.*, unpublished). The Brazilian genotypes were courtesy of the IAPAR centre in Parana and the Guyana accessions came from the CIRAD collection centre at Pointe Combi.

DNA extraction

Genomic DNA was extracted from ground leaves in accordance with an extraction method using a MATAB buffer and formerly described in (Cubry et al., 2008). The extracts were then purified using an anion-exchange resin column (NucleoBond AX20 from Macherey-Nagel, Düren, Germany).

Microsatellite markers

A set of 39 microsatellites markers was used to genotype the 293 individuals. A complete description of the markers used is given in Table 3.2. The markers originated from different sources. 28 were designed from microsatellite-enriched libraries of *C. canephora* or *C. arabica* (Poncet *et al*., 2004; Poncet *et al*., 2007; Combes *et al*., 2000). Four were from the Nestle Research Centre (Tours, France) and were designed on an EST library (Crouzillat, Nestlé, Tours, France, pers. comm.). Five were designed on the BAC-ends sequences of a BAC library of *C. canephora* (Leroy *et al.*, 2005) and 2 were designed on sequences from a sucrose synthase gene (Geromel *et al*., 2006; Cubry *et al*., 2008). The microsatellites were been mapped on an intraspecific *C. canephora* genetic map (Leroy, CIRAD, Montpellier,

France, pers. comm.) and were chosen to avoid redundancy in information and to enable complete coverage of the 11 linkage groups.

PCR and data acquisition

PCR reactions and data acquisition from fluorescently labelled PCR products were performed according to (Cubry et al., 2008). The data matrix was exported as a text file and formatted in Excel® software for CONVERT (Glaubitz, 2004) and CREATE (Coombs *et al.*, 2007) softwares, which were used to format the data for the different softwares used.

Genetic structure analysis

We used the main two groups of methods for investigating genetic structure (Pritchard *et al*., 2000a), i.e. a model-based method using a Bayesian implementation model with the Structure 2.2 program (Pritchard *et al*., 2000a; Falush *et al*., 2003) and a distance-based model using a dissimilarity matrix calculated with a simple-matching index and a Principle Coordinates Analysis (PCoA) (Perrier *et al.*, 2003) as implemented in the DARwin 5.0 software (Perrier and Jacquemoud-Collet, 2006). Those two methods were used simultaneously in order to better investigate the precise structure of our sample.

Structure runs were carried out on an high-performance computing server at Oslo University (Bioportal), with 200 000 burn-in repeats and 400 000 MCMC iterations for each K, 10 to 20 repeats per K. We used the transformation suggested by (Evanno *et al*., 2005) to help choose the optimum value for K (i.e. the uppermost level of structure assumed). The final choice of K was based on the four diagnostic plots of Evanno's transformation and the previously known structure. The parameters used for simulations were the admixture model with correlated allele frequencies as suggested by (Falush *et al*., 2003), all other parameters were set to default. Once the best K was chosen, we summarized the results of the n iterations of the model for the assumed K using CLUMPP software (Jakobsson and Rosenberg, 2006). Graphical displays of the results were made using Distruct software (Rosenberg, 2004). We assumed that an individual belonged to a group when its probability of ancestry was at least 0.80. Individuals not assigned to a cluster were mixed together in an admixed population.

For the whole dataset, we ran several runs of Structure with K varying from one to 25. To better investigate the precise structure of our data, we cross-checked the results obtained with the graphical methods and we ran Structure on a partitioned dataset in order to investigate lowest levels of structure, in relation to the results obtained with the graphical approach. For the partitioned datasets, K was allowed to vary from one to at least 9.

For a group identified in Structure in which we assumed a cryptic substructure, we produced a Neighbour-Joining (NJ) tree based on the dissimilarity matrix using DARwin. Population names were used to label the tree.

Basic statistics computation

Descriptive population genetics parameters were calculated, including Observed Heterozygosity (H_o), Gene Diversity (an analogue to Expected Heterozygosity (H_e)) and Mean Allele Number (N_a), for both the whole dataset and the diverse groups after reassignment. Confidence intervals were calculated using 1000 bootstraps. Computations were performed using PowerMarker 3.25 (Liu and Muse, 2005). Other descriptive statistics, including Stepwise Mutation Index (*SMMindex*, i.e. the percentage of alleles following the stepwise assumption per marker), or percentage of available data, were computed.
For the groups identified from the analysis, or for some populations, *F-statistics* (F_{is} and F_{st}) were computed using Arlequin, as well as an AMOVA (Analysis of Molecular Variance) which indicates the relative strength of each diversity level. For our study, we used analogues of classical F_{st} and derived distances by computing Slatkin's R_{st} which is more suited to microsatellites (Michalakis and Excoffier, 1996; Slatkin, 1995). Significance of the *F-statistics* was tested using 1000 permutations and a 5% threshold for the associate p-values. R_{st} analysis was used to assess the significance of the determined structure, as well as the differentiation between the considered populations. F_{is} and AMOVA provided information on the significance of the determined structure and the possibility of a resulting substructure.

Isolation by distance analysis and phylogeographical investigation

Since we used microsatellite markers, we had to consider their specificities in terms of evolution, in order to be able to put forward some hypotheses on the phylogeography of our species, as well as the demographic history of our populations. Microsatellites generally mostly fit a stepwise mutation model (SMM) (Goldstein and Pollock, 1997) and appropriate measures have been developed to take that specificity into account, including R_{st} (Slatkin, 1995) and $\delta\mu^2$ (Goldstein *et al.*, 1995a). Isolation by distance was investigated using Mantel tests between molecular distance matrixes derived from microsatellite data (matrixes of linearized R_{st} produced with the help of Arlequin) and geographical distance matrixes. Significance of the test was tested with 9999 permutations. We tested isolation by distance on a continent and country scale. Geographical distances were derived from survey information. Mantel tests were performed using GenALex software (Peakall and Smouse, 2006).

To investigate the phylogeny of our data in relation to geographical distribution, we computed the $\delta\mu^2$ distance, which has been shown to be linear with time (Goldstein *et al.*, 1995b). A WPGMA tree based on the distance matrix was constructed using DARwin to better understand relationships between the identified groups.

Results

Global genetic diversity and basic statistics

The results of the global analysis are given in Table 3.3. The whole sample showed a mean number of alleles per locus of 11.81 (462 alleles in total), ranging from 4 to 19. Gene Diversity, considered as informative of genetic diversity, had a mean of 0.72 per locus, ranging from 0.22 to 0.91. In contrast, the Observed Heterozygosity (H_o) was lower, ranging from 0.03 to 0.60, with a mean of 0.36. None of our markers appeared to be at Hardy-Weinberg (HW) equilibrium on the scale of the whole dataset.

SMMindex indicated that the great majority (i.e. 38 out of 39) of our microsatellite allele sizes tallied with a Stepwise Mutation Model. Only one (364) exhibited a significant proportion of alleles that did not follow the SMM model (data not shown) at a 5% threshold.

Genetic structure investigation

The results of the structure studies and graphical analyses on the different levels of investigation are provided in Figures 3.2 and 3.3.

The higher level of structure detected with the model-based method, and a comparison with the highest level of structure detected through graphical analysis. (Figures 3.2A and 3.3A)

The structure model including the admixture and correlated allele frequencies was run on the whole dataset of 293 individuals. The four Evanno diagnostic graphs suggested an uppermost structure level with two groups. Overall, those groups appeared to correspond to the previously described “Congolese” and “Guinean” groups (Berthaud, 1986). The individuals classified as admixed corresponded to cultivated genotypes or originated from populations collected near cultivated areas in Ivory Coast. The graphic representation strengthened that sample distribution but also highlighted the existence of a remaining structure in the defined groups (Figure 3.3 A). We then partitioned the initial dataset according to the probability of ancestry for the individuals, with an 80% threshold. All in all, the collection data (i.e. geographical origins) appeared to be robust on that level, with only one misclassified individual, G1036. Partitioned datasets regrouped 123 genotypes for the Congolese and 158 genotypes for the Guineans, 12 individuals were considered as admixed. The partitioned data, excluding the admixed group, were used separately for the following step.

Second level of structure runs and comparison with second-level graphical analysis. (Figures 3.2B and 3.3B)

A study of the structure results indicated 2 subgroups for each of the previously identified clusters. Graphical analysis confirmed those results with the first axis separating the indicated subgroups.

For the Guineans, Pelezi was separated from the other origins, while for the Congolese Nana was separated.

Nana and related genotypes had been already classed as the C diversity group. In contrast, strong evidence of a genetic structure in the Guinean group was a new result.

Four new partitioned datasets were constituted based on the CLUMPP summary of the structure simulation results at K=2 for both the Guinean and Congolese clusters. For the Guinean cluster, one group-population (Pelezi, 34 individuals) and one composite group (other Guinean, 122 individuals) were considered. For the Congolese cluster, one group-population (Nana, 43 individuals) and one composite group (other Congolese, 78 individuals) were adopted.

Admixed individuals (4 genotypes) were discarded from the following analysis and incorporated into the admixed composite population.

Third level of structure runs and comparison with third-level graphical analysis. (Figures 3.2C and 3.3C)

For the two group-populations highlighted in the above analysis (i.e. Nana and Pelezi), structure failed to converge to a clear maximum likelihood and graphical analysis did not show a possible substructure.

For the Other Guinean group, Evanno's transformation gave a weak signal at K=2. When using this clustering to highlight individuals in the factorial analysis, no clear separation could be established between those two groups. According to previous studies based on isozymes, along with the potential for seed dispersal and the biology of our species (strictly allogamous), we assumed that there was no clear but a fine separation between populations in this group, like the results obtained on wild olive populations by Belaj *et al*., (2007).

For the Other Congolese group, Evanno's diagnostic plots give a strong signal at K=3. This clustering was confirmed by the factorial analysis and corresponded to 3 previously described groups using RFLP (SG1, SG2 and B).

In brief, we found a structure with four group-populations (groups composed by one population: Pelezi, Libengé, Niaouli and Nana) and two composite groups (groups with different populations: SG2 and Other Guinean).

Genetic diversity and differentiation parameters of the inferred diversity clusters

The results of diversity analyses on different levels are given in Tables 3.4 and 3.5. *F-statistics* analysis (R_{st}) and AMOVA (R_{st}-based) results are given in Tables 3.6 and 3.7. R_{is} per population are indicated in Table 3.8. F_{st} and F_{st}-based AMOVAs are provided as supplemental material.

Diversity on the first structure level

The mean number of alleles was 10.21 for the Congolese group and 7.44 for the Guinean group. The mean expected heterozygosity ranged from 0.53 for the Guinean group to 0.73 for the Congolese group. As for the whole dataset, the observed heterozigosity was lower than gene diversity, with 0.30 and 0.40 for the Guinean and Congolese groups respectively. Comparisons between groups led to the assumption of greater diversity in the Congolese group than in the Guinean group.

The *F-statistics* indicated a clear differentiation between the two groups, with significant values at a 5% level. However, R_{is} appeared to be high with 0.30 (0.25 and 0.35 for the Guinean and Congolese groups respectively), indicating a residue of structure within those groups.

Diversity on the second structure level

On this level we were able to compare diversity in two group-populations and two composite groups. For Pelezi and Nana, the group-populations, the mean allele number was 3.08 and 4.74 respectively, which was significantly lower than the values obtained for Other Guinean and Other Congolese (6.90 and 8.96 respectively). In terms of gene diversity, only the Other Congolese group appeared to have a significantly higher value than the other populations (giving the 95% confidence interval).

AMOVAs and *F-statistics* exhibited high values which were all significant at a 5% threshold. More than 40 percent of the variation appeared to originate from groups and populations. That was greater than the percentage explained by the two clusters of the first level. A substantial residue of the variation was still contained at the within-individual level. R_{is} values appeared to be lower than for the first level analysis, ranging from 0.08 to 0.26. Two clusters had very low R_{is} values, Nana (0.08) and Pelezi (0.09), whereas the other 2 had higher values. Those results validated the absence of structure within Pelezi and Nana, as expected from structure and graphical Analysis and indicated a residue within Other Congolese and Other Guinean.

Diversity on the third structure level

Niaouli and Libengé appeared to be comparable for several diversity parameters. The values obtained for gene diversity were similar to those obtained for Nana, Pelezi and Other Guinean, indicating a comparable amount of diversity within those clusters. For the SG2 cluster, it appeared that the diversity within that group was considerably higher than for the other groups.

The AMOVA analysis indicated only 5.4% of the variation occurring among individuals within the defined populations and more than 55% due to group or population subdivision. The R_{is} values were mostly low with an overall value of 0.11. Only 2 clusters, Niaouli and Other Guinean, exhibited a value greater than 0.14, showing the limited structure among the other populations and a putative cryptic structure within those 2 clusters. However, the value obtained for Niaouli was not very robust since it was based on only 8 genotypes. In order to test those results, we proceeded with a closer investigation of the structure within the Other Guinean group.

Diversity and fine structure of the Other Guinean cluster

We investigated the diversity within this cluster by drawing up a Neighbour-Joining tree based on a dissimilarity matrix. We used population information for the genotypes to label the tree (Figure 3.4).

The tree structure mainly corresponded to the different origins within the Other Guinean group. Mouniandougou and Pine genotypes were mixed up in the tree, even though some Pine genotypes were more distant. The Ira 1 and Ira 2 populations also appeared closely related. The populations were globally well defined but distances were small and the internal branches of the tree were very short. The populations represented a sort of continuum of diversity, with a composite population, the Cultivated Ivorian, which appeared to recover almost all that continuum. We chose to keep the population names to test for isolation by distance on an Ivory Coast and continent scale.

We tested the subdivision into populations by computing an AMOVA and pairwise R_{st} between populations in this group. Populations with only one individual (i.e. Sabregue and Unknown) were discarded from this analysis, with the smallest population analysed comprising 8 individuals. All the pairwise R_{st} values appeared to be significant but were lower than those previously obtained. Moreover, the AMOVA clearly indicated that the population level was responsible for about 15% of variation (significant value at a 5% threshold), the lowest value obtained throughout the study.

Isolation by distance

Isolation by distance was tested using 11 populations, 5 resulting from structure analysis (Niaouli, SG2, Libengé, Nana and Pelezi) and 6 from origin information in the Other Guinean cluster. Geographical distances were calculated on the basis of geographical coordinates of the populations for the Guinean, the Nana and the Libengé genotypes. For SG2, the putative origin of the cultivated genotypes from INEAC, Yangambi, was used. Cultivated Ivorian and admixed individuals were discarded from this analysis. A country (Ivory Coast) and a continent scale analysis were performed. No clear evidence of isolation was detected on the country level (R = -0.232, associate p-value = 0.304) but a significant signal was detected on the continent-scale (R = 0.749, associate p-value = 0.004). Log transformation of the geographical distances did not significantly change the results (data not shown).

Phylogeography

In order to investigate relationships between groups, we used the 6 clusters defined by our structure analysis. The resulting tree is shown in Figure 3.5. In brief, we found a separation between the Guinean and Congolese groups in the first place, followed by more recent divergence within those 2 groups. For the Guinean group, the Pelezi versus other populations separation appeared to be a recent event, while divergences among the Congolese group should be more ancient. For the Congolese groups, Libengé and SG2 appeared to be the last groups to be divided, Nana was separate a short time before and the first divergent group was Niaouli. This tree could be superposed over the geographical distribution of our sampling origins.

Discussion

Genetic structure investigation

The comparison between the structure simulations study and graphical analysis was of great use in assessing the diversity structure of our species. Even though our nested-structure analysis performed well in detecting global clusters of diversity, it failed to converge to a very fine-scale structure on a population level in the Guinean group. That population structure has been validated elsewhere with graphical analyses. The UW group identified by (Musoli *et al*., unpublished) as a separate group was classified here with the SG2, highlighting the proximity between those groups and the difficulty in clearly differentiating with a small sample (only 3 genotypes in our study).

The Other Guinean group was previously studied on an Ivory Coast level using isozymes by (Berthaud, 1986). His study concluded in the absence of a geographical structure among the Ivorian populations, but no tests were carried out. The results obtained in our study showed that there was weak differentiation between populations. Seeds can be dispersed over a long distance and should explain the existence of relay-trees or contamination of a population by allogenous germplasm, as well as the establishment of new populations. The isolation by distance analysis indicated an absence of correlation between geographical and genetic distances on the scale of 100 km, while there was a correlation on a continent scale (this study) and on a forest scale in Uganda (scale of a few km, Musoli *et al*., unpublished). The genetic landscape of *Coffea canephora* in Ivory Coast appears not to be fixed with inter-population exchanges based on both pollen and seed dispersal, as well as possible migration events between wild and cultivated compartments, as shown by the existence of hybrids in the cultivated fields.

Origin of genetic structure: the Last Glacial Maximum (LGM) refuges hypothesis

Our sampling of one origin per previously known group, except for the Guinean group and for SG2, did not allow us to clearly demonstrate that there was a real structure in the diversity groups of our species. However, some other indications strengthened this hypothesis. A study of another population from the Luki region of RDC (near the Niaouli putative geographical origin) indicated that spontaneous genotypes from that region were well clustered with the Niaouli accessions. That study confirmed the reality of an SG1 diversity group, since those genotypes came from collections of spontaneous material (rapport final INCO-COWIDI, 2007).

For the C group, previous studies incorporating some Nana genotypes and other origins, based on isozyme data, showed that those genotypes were genetically closest to Cameroonian accessions than to genotypes from Libengé (B group) (Berthaud, 1986; Dussert *et al*., 1999), despite their geographical proximity. The existence of 2 different origins might explain that divergence. The same event occurred for the differentiation between SG1 and SG2. (Berthaud, 1986) put forward the hypothesis that the origin of those different groups was the existence of glacial refuges. The Guinean and Congolese groups, the first to have diverged, may correspond to the firsts steps of the reduction in the habitat area of the last glaciation, and are still separated today with an absence of spontaneous *C. canephora* trees in the Togo-Benin region. This structure has also been reported for some other coffee trees, such as *Coffea liberica* (Berthaud, 1986) and some other species belonging to other genera e.g. *Cola nitida* (Sie *et al.*, 2005). This separation corresponds well to two major refuge areas in the Last Glacial Maximum (LGM, 18000 years BP) as described by (White, 1979). The indications of (Maley, 1996) and (Adams and Faure, 1997) give several potential refuges in the central Africa region that tally with our structure. One of the refuges of the Atlantic coast of the Gulf of Guinea in the region of the current Gabon may correspond to SG1, while a refuge on Mount Cameroon may have given rise to the C group. SG2 and B should originate from a unique refuge area, since they were the less divergent in the Congolese group according to the $\delta\mu^2$ distance and they both originated from the Congo Basin. The divergence of SG2 from B might also be due to a selection process since that material only consisted of improved material.

The same question can be raised regarding genetic differentiation within the Guinean group. We highlighted high genetic differentiation of the Pelezi population while geographical distance was low. That population was surveyed on the border of the *Coffea canephora* area and was most probably subjected to strong ecological constraints, such as drought stress. No other populations of coffee trees were observed within a radius of at least fifty km (T Leroy, CIRAD, Montpellier, France, pers. comm.), reinforcing the idea of the isolation of that population and its originating from abiotic constraints.

Brazilian, Nemaya and Ugandan origins

The study of some accessions cultivated in South America, as well as cultivated Ugandan material, revealed the limited diversity of the sampled material. There is some evidence that *Coffea canephora* growing worldwide is mainly based on two genetic groups (SG2, as mentioned in our study, and SG1). The narrow genetic background of material

currently cultivated opens up the way for an improvement of the crop with new selected material. We also highlighted the SG2 origin of the Nemaya genotypes from the CATIE collection used as rootstocks for nematode resistance.

Identification of some target populations to perform Linkage Disequilibrium analyses

One of the main goals of this analysis was to determine some populations with limited structure to launch Linkage Disequilibrium (LD) studies avoiding detection of false positives due to the genetic structure, which has been reported to be one of the major limitations of such approaches (Pritchard *et al*., 2000b; Yu *et al*., 2006). We showed that the 6 clusters resulting from the structure analysis had low inbreeding coefficients and that genetic diversity mainly came from the population level. We propose carrying out LD studies in 2 group-populations (i.e. groups consisting of only one population), Nana and Pelezi, and in 2 composite-groups, SG2 and Other Guinean. Moreover some resampling will be carried out in both the Congolese and Guinean groups to build core-collections for each group. By comparing LD patterns in those different samples covering most of the species' diversity it will be possible to determine the potential of association studies in *C. canephora*.

Conclusion and prospects

This study encompasses the largest sampling operation in terms of molecular markers and genotypes undertaken on the *Coffea canephora* species. We confirmed the previously hypothesised genetic structure of this species and added a new substructure to one major group. However, for a more effective investigation of *Coffea canephora* structure, some sampling is required in DRC, as well as in Cameroon, Gabon and South-East Ivory Coast. It will improve our ability to determine the relevance of the groups defined in this study. Phenotypic characterisation of the different diversity clusters found in this study will be carried out in order to find some interesting potentials for drought tolerance, and for quality and productivity improvement in the cultivated *C. canephora* compartment.

We evaluated the genetic resources that can be found in collections worldwide for *Coffea canephora*. The genetic origin of the cultivated material, mainly based on 2 diversity groups (SG1 and SG2), suggested that the genetic bases of the cultivated material could be substantially improved. A new group-population (Pelezi) was clearly identified and its geographical location, as well as the environmental stresses it has undergone, indicates a potential source of diversity for drought tolerance. A model-based analysis of structure performed up to the “diversity groups” level, but failed in identifying local populations,

indicating the limitations of this approach for studying and demonstrating structure on a very fine level in our species. The hypothesis of a genetic structure drawn by the last glacial events first suggested by (Berthaud, 1986) was reinforced by our analysis.

This work provides the basis for future molecular characterisation of the Divo reference collection with a subset of the markers used, like the study undertaken on *Vitis vinifera* (Le Cunff *et al.*, 2008). It also provides the ability to identify new material from other collections, like the study carried out on Luki genotypes. Such characterisation will help standardise the collection and will provide opportunities to build core-collections to reduce the cost of maintenance and phenotypic studies in our perennial species maintained only in the field.

This study was part of a global study on Linkage Disequilibrium (LD) in *Coffea canephora* and enabled the identification of natural and composite populations with a controlled structure that are easily usable for such applications. LD studies, which will be performed on those different samples, will enable the identification of association populations for different purposes, giving the LD pattern they will exhibit, including whole genome scans for QTL identification or candidate gene approaches.

Acknowledgements

We thank the Nestlé Research Centre in Tours (France) for providing us with some EST derived microsatellite markers. The main plant material came from the Divo Research Centre, CNRA, Ivory Coast, and from the Naro-CORI collection in Uganda. Other materials were kindly provided by IAPAR, Paranà state, Brazil and CATIE, Costa Rica.

LITERATURE CITED

Adams JM, Faure H. 1997. Preliminary vegetation maps of the world since the Last Glacial Maximum: an aid to archaeological understanding. *Journal of Archaeological Science,* **24**: 623-647.

Adibolo Y, Bertrand B. 1988. Etude de l'origine de la variété Niaouli au Togo et au Bénin. *Café Cacao Thé,* **32**: 293-297.

Belaj A, Muñoz-Diez C, Baldoni L, Porceddu A, Barranco D, Satovic Z. 2007. Genetic diversity and population structure of wild olives from the North-Western Mediterranean assessed by SSR markers. *Ann Bot (Lond),* **100**: 449-58.

Berthaud J. 1986. *Les ressources génétiques pour l'amélioration des caféiers africains diploïdes,* Paris, ORSTOM.

Berthaud J, Anthony F, Le Pierrès D. 1984. Les caféiers de la Nana. Résultats des observations faites en collection en Côte d'Ivoire. *Café Cacao Thé,* **28**: 3-13.

Berthaud J, Guillaumet J-L. 1978. Les caféiers sauvages en Centrafrique. Résultats d'une mission de prospection (janvier-février 1975). *Café Cacao Thé,* **22**: 171-187.

Bioportal. a High-Performance Computing server, available at http://www.bioportal.uio.no/. Oslo, University of Oslo.

Bodard L. 1965. Historique des caféiers de RCI.

Combes M, Andrzejewski S, Anthony F, Bertrand B, Rovelli P, Graziosi G, Lashermes P. 2000. Characterization of microsatellite loci in Coffea arabica and related coffee species. *Mol Ecol,* **9**: 1178-80.

Coombs JA, Letcher BH, Nislow KH. 2007. CREATE 1.0 – Software to create and convert codominant molecular data. Available online at http://www.lsc.usgs.gov/CAFL/Ecology/Software.html.

Cubry P, Musoli P, Legnaté H, Pot D, de Bellis F, Poncet V, Anthony F, Dufour M, Leroy T. 2008. Diversity in coffee assessed with SSR markers: structure of the genus Coffea and perspectives for breeding. *Genome,* **51**: 50-63.

Davis A, Stoffelen P. 2006. An annotated taxonomic conspectus of the genus Coffea (Rubiaceae). *Botanical Journal of the Linnean Society,* **152**: 465-512.

Dussert S, Lashermes P, Anthony F, Montagnon C, Trouslot P, Combes MC, Berthaud J, Noirot M, Hamon S. 1999. Le caféier, Coffea canephora. In: Hamon P, Seguin M, Perrier X, Glaszmann JC eds. *Diversité génétique des plantes tropicales cultivées.* Montpellier, CIRAD.

Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol,* **14**: 2611-20.

Falush D, Stephens M, Pritchard J. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics,* **164**: 1567-87.

Geromel C, Ferreira L, Guerreiro S, Cavalari A, Pot D, Pereira L, Leroy T, Vieira L, Mazzafera P, Marraccini P. 2006. Biochemical and genomic analysis of sucrose metabolism during coffee (Coffea arabica) fruit development. *J Exp Bot,* **57**: 3243-58.

Glaubitz J. 2004. CONVERT: a user friendly program to reformat diploid genotypic data for commonly used population genetic software packages. *Molecular Ecology Notes,* **4**: 309-310.
Goldstein D, Pollock D. 1997. Launching microsatellites: a review of mutation processes and methods of phylogenetic interference. *J Hered,* **88**: 335-42.
Goldstein D, Ruiz Linares A, Cavalli-Sforza L, Feldman M. 1995a. An evaluation of genetic distances for use with microsatellite loci. *Genetics,* **139**: 463-71.
Goldstein D, Ruiz Linares A, Cavalli-Sforza L, Feldman M. 1995b. Genetic absolute dating based on microsatellites and the origin of modern humans. *Proc Natl Acad Sci U S A,* **92**: 6723-7.
Jakobsson M, Rosenberg NA. 2006. CLUMPP : CLUster Matching and Permutation Program. http://rosenberglab.bioinformatics.med.umich.edu/software.html.
Le Cunff L, Fournier-Level A, Laucou V, Vezzulli S, Lacombe T, Adam-Blondon A, Boursiquot J, This P. 2008. Construction of nested genetic core collections to optimize the exploitation of natural diversity in Vitis vinifera L. subsp. sativa. *BMC Plant Biol,* **8**: 31.
Leroy T, Marraccini P, Dufour M, Montagnon C, Lashermes P, Sabau X, Ferreira L, Jourdan I, Pot D, Andrade A, Glaszmann J, Vieira L, Piffanelli P. 2005. Construction and characterization of a Coffea canephora BAC library to study the organization of sucrose biosynthesis genes. *Theor Appl Genet,* **111**: 1032-41.
Liu K, Muse S. 2005. PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics,* **21**: 2128-9.
Maley J. 1996. The African rain forest - main characteristics of changes in vegetation and climate from the Upper Cretaceous to the Quaternary. In: Alexander IJ, Swaine MD, Watling R eds. *Proceedings of the Royal Society of Edinburgh.* Edinburgh, The Royal Society of Edinburgh.
Michalakis Y, Excoffier L. 1996. A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. *Genetics,* **142**: 1061-4.
Montagnon C. 2000. *Optimisation des gains génétiques dans le schéma de sélection récurrente réciproque de Coffea canephora Pierre,* PhD, Ecole Nationale Supérieure Agronomique de Montpellier, Montpellier.
Montagnon C, Leroy T, Yapo A. 1992. Diversité génotypique et phénoytpique de quelques groupes de caféiers (Coffea canephora Pierre) en collection. *Café Cacao Thé,* **36**: 187-198.
Montagnon C, Leroy T, Yapo A. 1993. Caractérisation et évaluation de caféiers Coffea canephora prospectés dans des plantations de Côte d'Ivoire. *Café Cacao Thé,* **37**: 115-119.
Musoli P, Cubry P, Aluka P, Billot C, Dufour M, De Bellis F, Kyetere D, Ochugo J, Bieysse D, Charrier A, Leroy T. unpublished. A new genetic group from Uganda within Coffea canephora Pierre.
Neale D. 2007. Genomics to tree breeding and forest health. *Curr Opin Genet Dev,* **17**: 539-44.
Peakall R, Smouse PE. 2006. genalex 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes,* **6**: 288-295.

Perrier X, Flori A, Bonnot F. 2003. Data analysis methods. In: Hamon P, Seguin M, Perrier X, Glaszmann JC eds. *Genetic diversity of cultivated tropical plants*. Montpellier, Enfield, Science Publishers.

Perrier X, Jacquemoud-Collet JP. 2006. DARwin software. http://darwin.cirad.fr/darwin.

Poncet V, Dufour M, Hamon P, Hamon S, de Kochko A, Leroy T. 2007. Development of genomic microsatellite markers in Coffea canephora and their transferability to other coffee species. *Genome,* **50**: 1156-61.

Poncet V, Hamon P, Minier J, Carasco C, Hamon S, Noirot M. 2004. SSR cross-amplification and variation within coffee trees (Coffea spp.). *Genome,* **47**: 1071-81.

Portères R. 1937. Etude sur les caféiers spontanés de la section Eucoffea. Leur répartition, leur habitat, leur mise en culture et leur sélection en Côte d'Ivoire. I Répartition et habitat. *Annales de l'Afrique Occidental Française et Etrangère,* **1**: 68-91.

Portères R. 1959. Valeur agronomique des caféiers de types Kouilou et Robusta cultivés en Côte d'Ivoire. *Café Cacao Thé,* **3**: 3-14.

Pritchard J, Stephens M, Donnelly P. 2000a. Inference of population structure using multilocus genotype data. *Genetics,* **155**: 945-59.

Pritchard J, Stephens M, Rosenberg N, Donnelly P. 2000b. Association mapping in structured populations. *Am J Hum Genet,* **67**: 170-81.

Rosenberg NA. 2004. DISTRUCT : a program for the graphical display of population structure. Available at http://rosenberglab.bioinformatics.med.umich.edu/software.html. *Molecular Ecology Notes,* **4**: 137-138.

Sie RS, N Goran JAK, Montagnon C, Akaffou DS, Cilas C. 2005. Assessing genetic diversity in a germplasm collection of kola trees (Cola nitida (Vent.) Schott and Endl.) using enzymatic markers. *Plant Genetic Resssources Newsletter*: 59-64.

Slatkin M. 1995. A measure of population subdivision based on microsatellite allele frequencies. *Genetics,* **139**: 457-62.

Thomas AS. 1935. Types of Robusta coffee and their selection in Uganda. *The East African Agricultural Journal,* **1**: 193-198.

Tuberosa R, Salvi S. 2006. Genomics-based approaches to improve drought tolerance of crops. *Trends Plant Sci,* **11**: 405-12.

White F. 1979. The Guineo-Congolian region and its relationships to other phytochoria. *Bulletin du Jardin Botanique National de la Belgique,* **49**: 11-55.

Yu J, Pressoir G, Briggs W, Vroh Bi I, Yamasaki M, Doebley J, McMullen M, Gaut B, Nielsen D, Holland J, Kresovich S, Buckler E. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet,* **38**: 203-8.

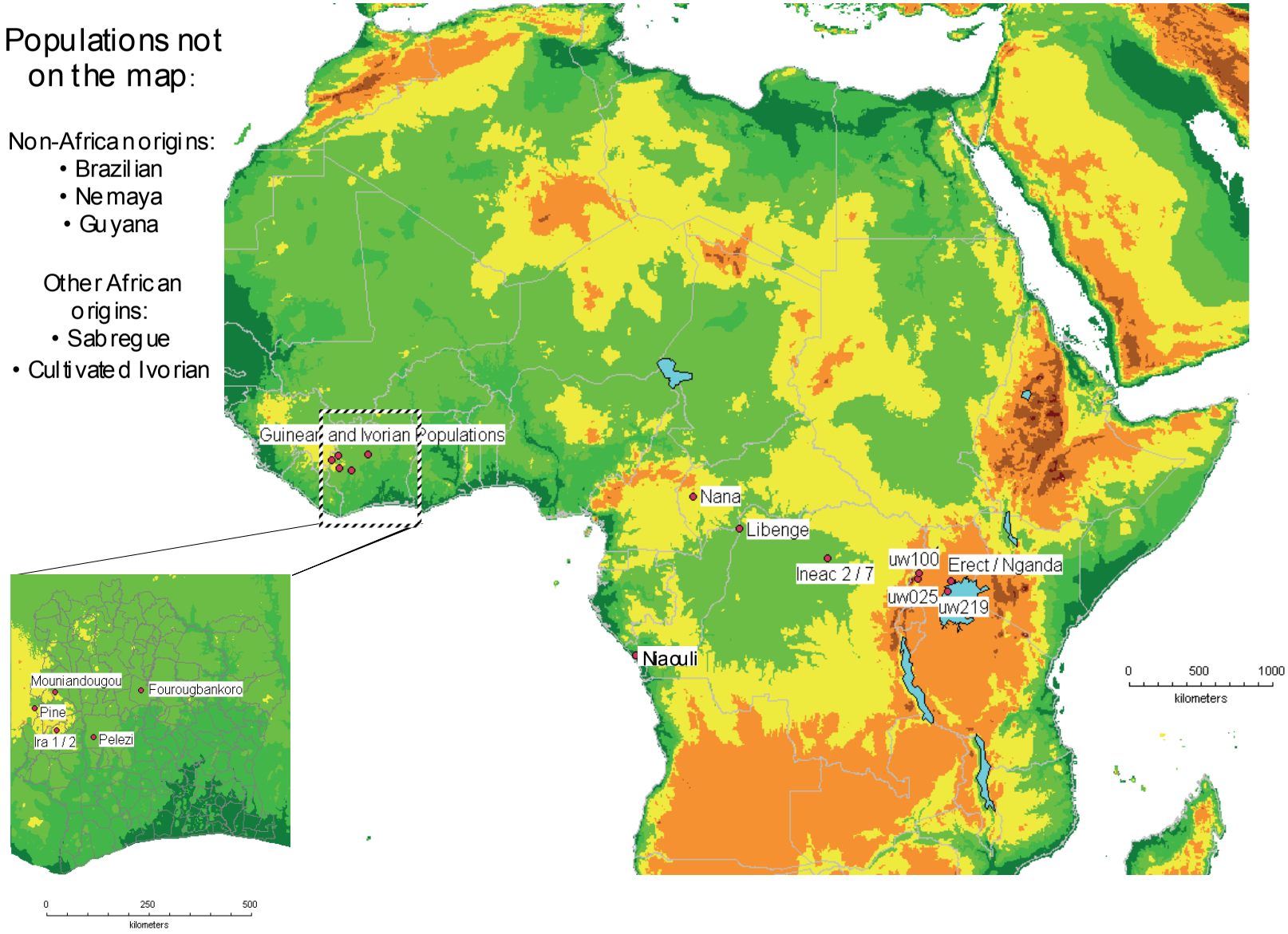


Figure 3.1: Location of the sampled populations throughout Africa. Sampling complements were obtained from collections worldwide.

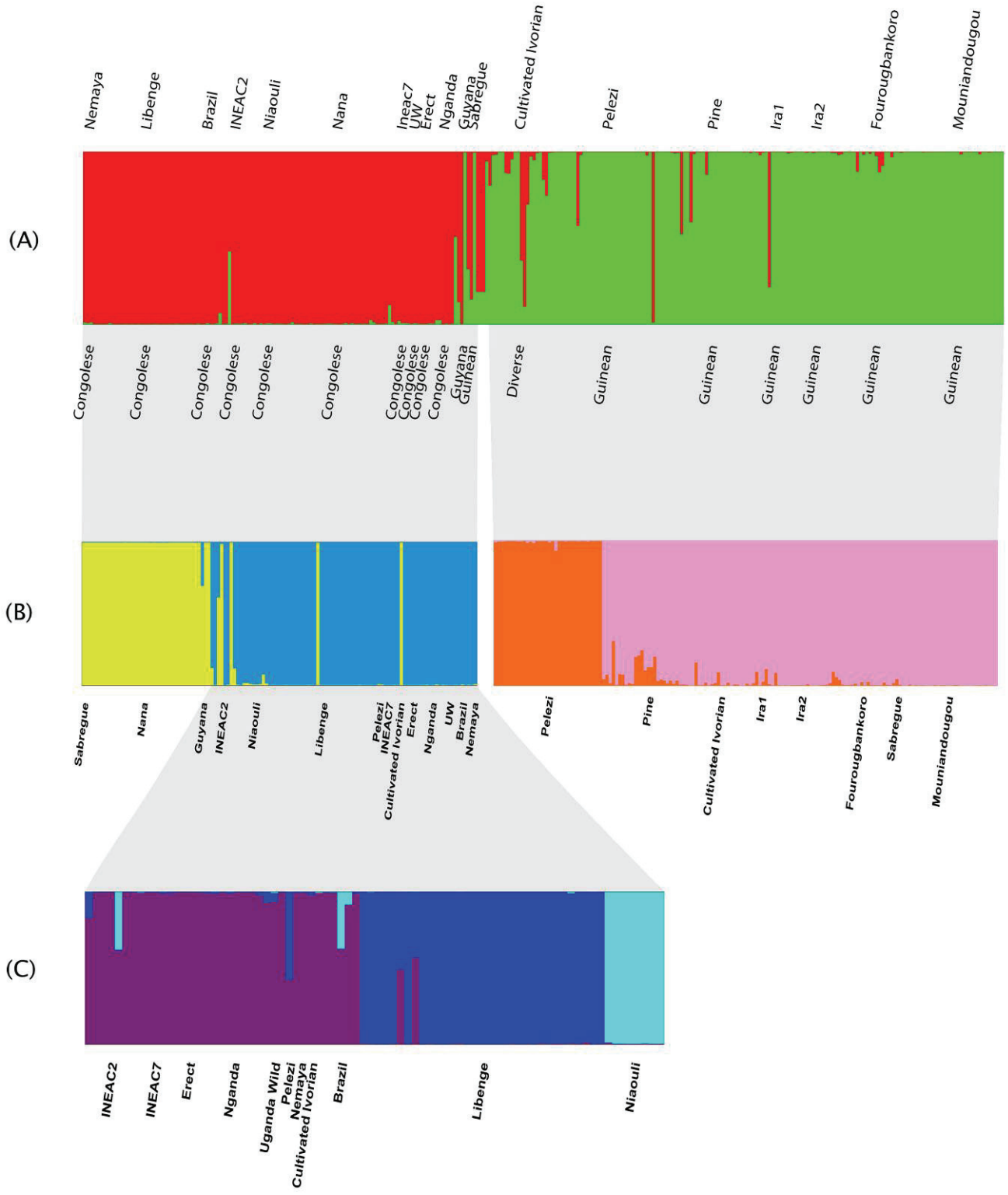


Figure 3.2: Structure outputs on different levels of structure investigation. (A) = Structure output for the whole sample, K=2, Congolese group (red) and Guinean group (green). Lower labels are for the putative origin of the populations, upper labels are the populations names. (B) = Structure outputs for the two datasets (Congolese on the left, Guinean on the right) of the 2nd level of investigation. Yellow = C, Blue = Other Congolese, Orange = Pelezi, Pink = Other Guinean (C) = Structure output for the "Other Congolese" group for the 3rd level of investigation. Purple = SG2, Dark-Blue = B, Light-Blue = SG1.

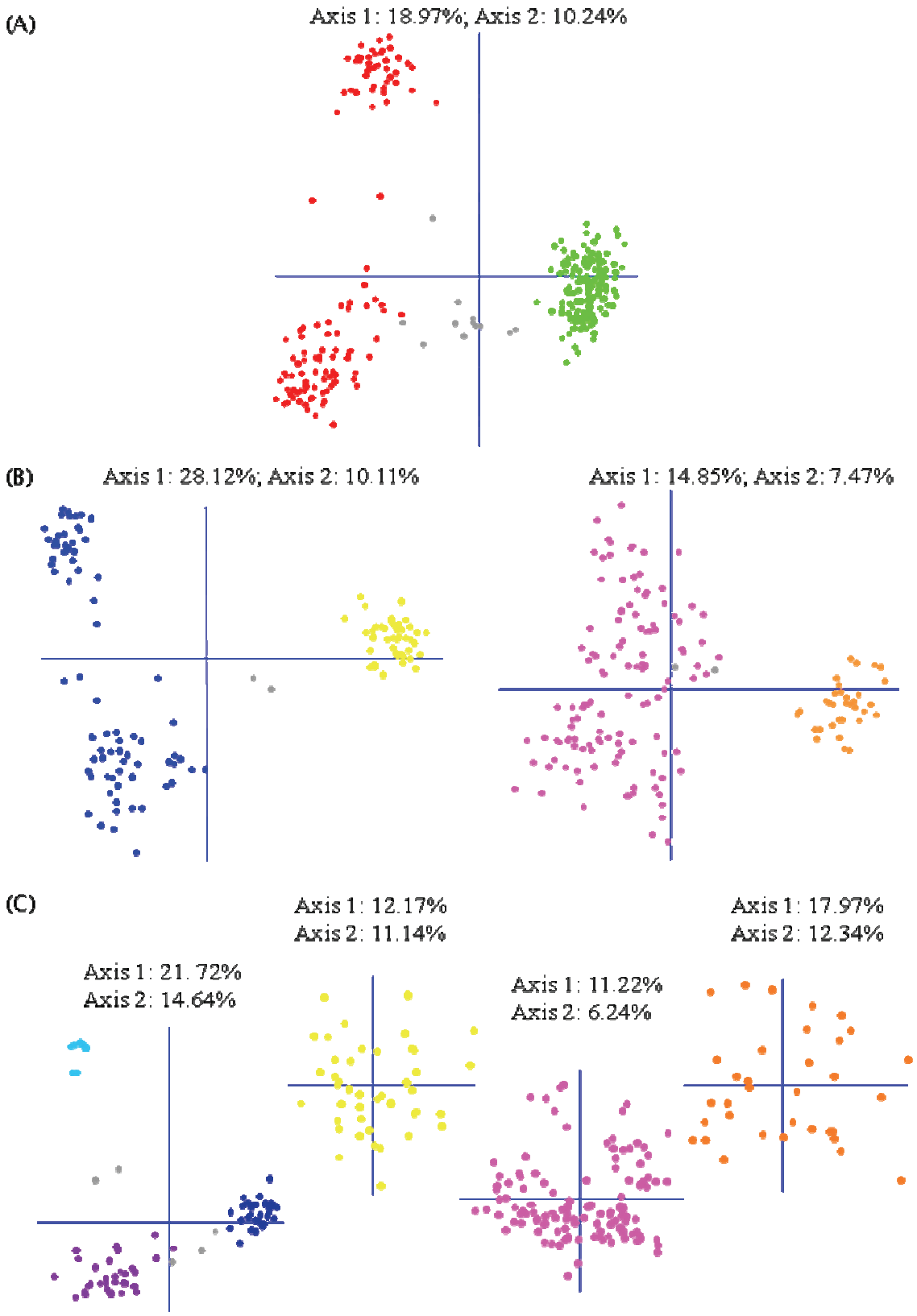


Figure 3.3: Factorial Analysis based on a Dissimilarity matrix for the different levels of structure investigation. (A) = Factorial analysis output for the whole sample, Congolese group (red) and Guinean group (green). (B) = Factorial analysis outputs for the two datasets (Congolese on the left, Guinean on the right) of the 2nd level of investigation. Yellow = C, Blue = Other Congolese, Orange = Pelezi, Pink = Other Guinean. (C) = Factorial analysis outputs for the 3rd level of investigation. "Other Congolese" group on the left, followed by the C group, the Other Guinean group and the Pelezi group. Purple = SG2, Dark-Blue = B, Light-Blue = SG1 Yellow = C, Blue = Other Congolese, Orange = Pelezi, Pink = Other Guinean.

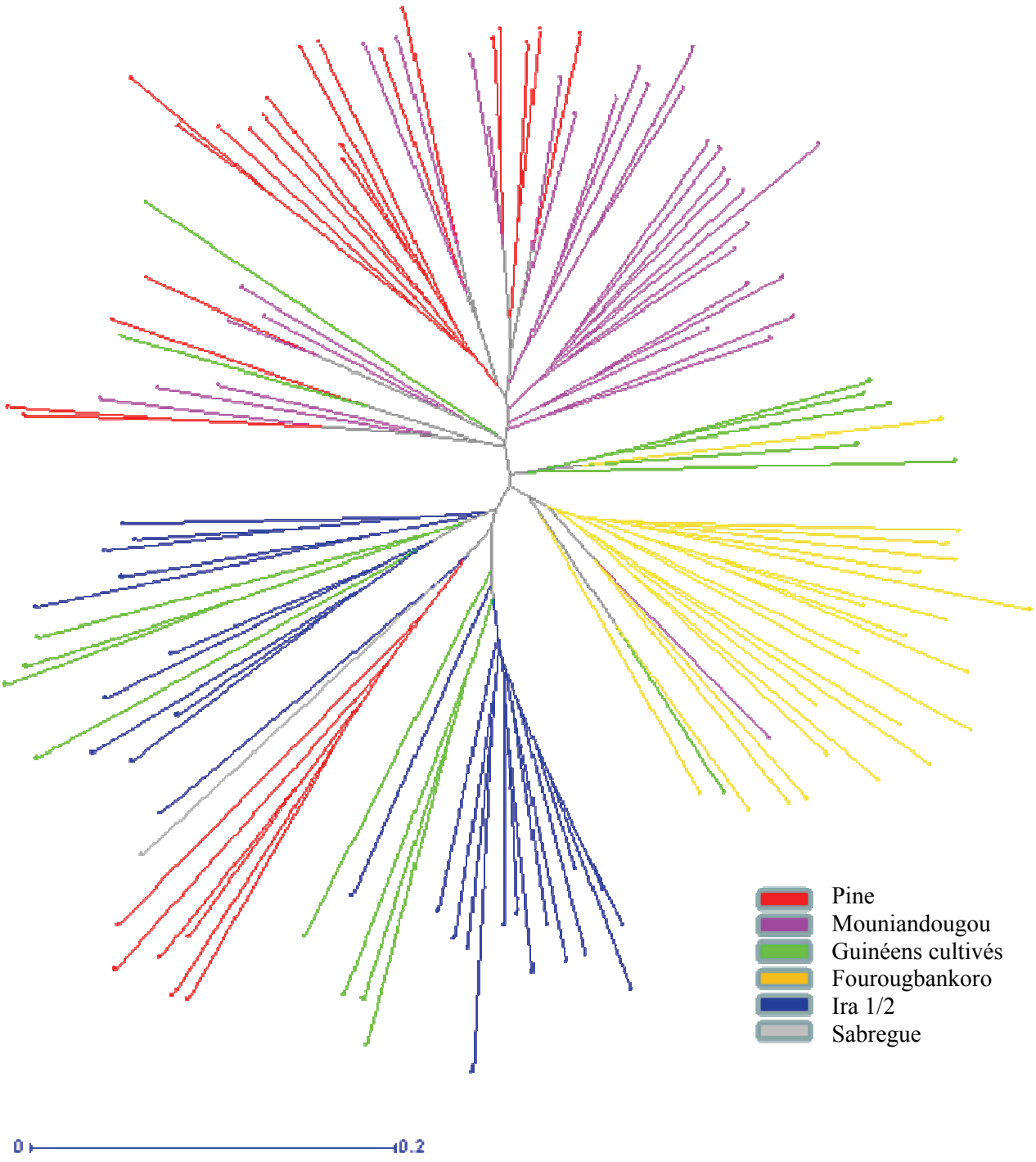


Figure 3.4: Neighbour-Joining tree based on a Dissimilarity matrix for the "Other Guinean" group. The legend indicates the colour-population correspondance.

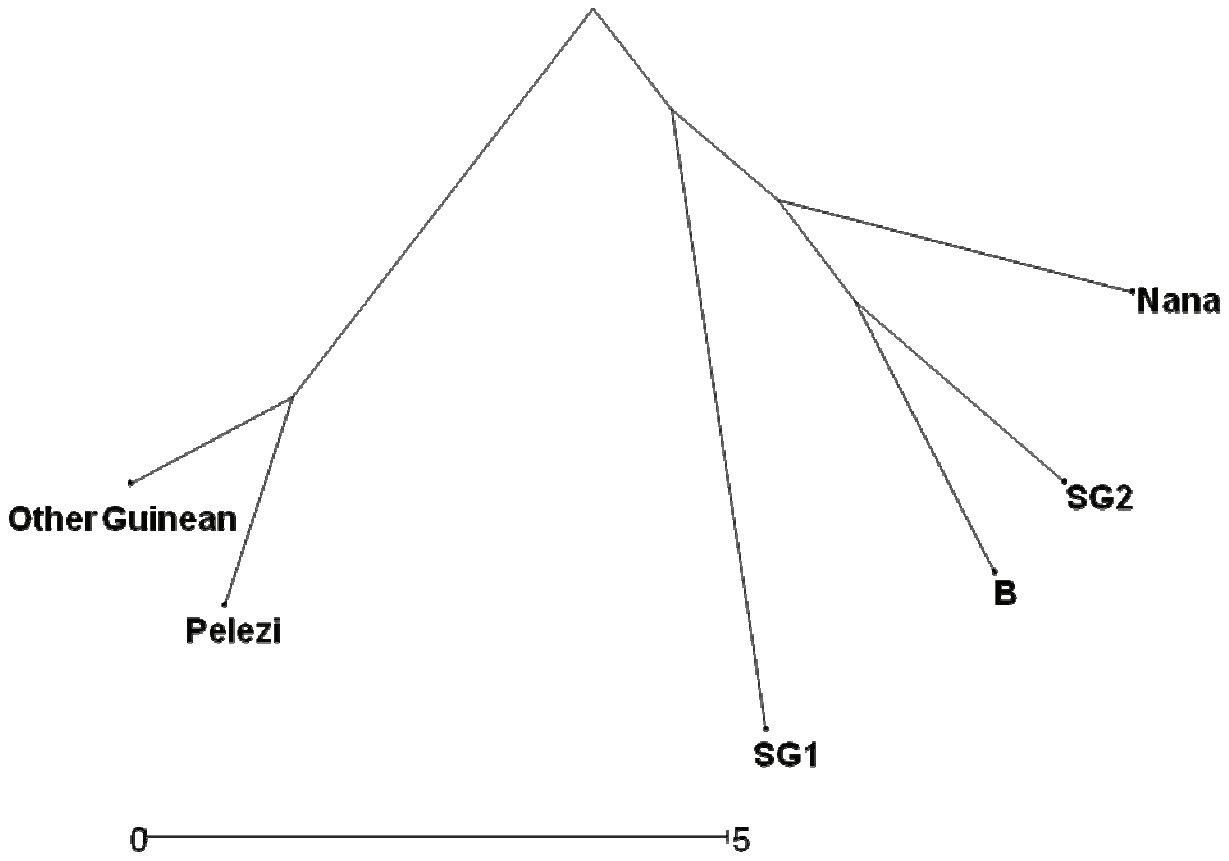


Figure 3.5: WPGMA tree based on $\delta\mu^2$ distance between the 6 identified diversity clusters.

Table 3.1: List of sampled populations used in this study with provider and putative origin

Population	Number of genotypes	Putative group based on collection data	Collection or Provider	Putative Origin	Type
Nemaya	2		CATIE	DRC	unknown
Brazil	5		Brasil	unknown	unknown
Libengue	34	B	CNRA – Divo	CAR near Libengé	wild
INEAC2	13	SG2	CNRA – Divo	INEAC (Yangambi)	improved
Niaouli	8	SG1	CNRA – Divo	Gabon or Congo via Benin	cultivated
Nana	36	C	CNRA – Divo	CAR	wild
INEAC7	5	SG2	CNRA – Divo	INEAC (Yangambi)	improved
Pelezi	36	G	CNRA – Divo	Ivory Coast	wild
Pine	31	G	CNRA – Divo	Guinea	wild
Cultivated Ivorians	24	G, SG2, Hybrids	CNRA – Divo	Ivory Coast	cultivated
IRA1	9	G	CNRA – Divo	Ivory Coast	wild
IRA2	17	G	CNRA – Divo	Ivory Coast	wild
Fourougbankoro	21	G	CNRA – Divo	Ivory Coast	wild
Mouniandougou	31	G	CNRA – Divo	Ivory Coast	wild
Sabregue	2	G	CNRA – Divo	Ivory Coast	spontaneous
Guy	4	Hybrids	CIRAD - French Guyana	unknown	unknown
Erect	4	SG2	Naro-CORI	Uganda	cultivated/improved
Nganda	8	SG2	Naro-CORI	Uganda	cultivated/improved
UW	3	UW	Naro-CORI	Uganda	wild

Table 3.2: List of the 39 microsatellite markers used in this study

Marker Name	EMBL acc number	Motif	Primer F	Primer R	Sequence origin	Primer origin
BES2M2	confidential	confidential	confidential	confidential	Leroy, 2005	Cubry, 2008 (unpub)
257	AJ250257	(ca)9	GACCATTACATTTCACACAC	GCATTTTGTTGCACACTGTA	Combes, 2000	Poncet, 2004
309	AM408738	(ac)9	AGCAACATTTCCCAGTCAA	GACCGCAATTTTCTTGTTTC	Dufour, 2001	Poncet, 2007
313	AM408741	(tg)9(gcat)3(cata)3	CGTGTGTGAGATAAAAATACAAG	CCGAAAACATTGCTACAGA	Dufour, 2001	Poncet, 2007
328	AM408753	(at)6(tg)8	CACCTTTTGAGTTTTGAGTTGG	AAAATAAACCCCCTTCGTTCA	Dufour, 2001	Poncet, 2007
341	AM231549	(ca)13	CATTGGTGTCAAGGGTCAAG	AAAGTATCAGAAGGAAAAGTCTCGTAA	Dufour, 2001	Poncet, 2007
355	AM231552	(tg)15	CTATGATGTCTTCCAACCTTCTAAC	GGTCCAATTCTGTTTCAATTTC	Dufour, 2001	Poncet, 2007
356	AM231553	(tg)14	TGAAGTCAACCTGAATACCAGA	ACGCACGCACGAATG	Dufour, 2001	Poncet, 2007
364	AM231556	(a)21	AGAAGAATGAAGACGAAACACA	TAACGCCTGCCATCG	Dufour, 2001	Poncet, 2007
367	AM231557	(ac)12	TCAATCCCTGTATTCCTGTTT	CTAGGCACTTAAAATCTCTATAACG	Dufour, 2001	Poncet, 2007
368	AM231558	(tg)13	CACATCTCCATCCATAACCATTT	TCCTACCTACTTGCCTGTGCT	Dufour, 2001	Poncet, 2007
384	AM231560	(ac)10	ACGCTATGACAAGGCAATGA	TGCAGTAGTTTCACCCTTTATCC	Dufour, 2001	Poncet, 2007
392	AM231562	(tc)16	AAGGTATTGGTCTGCCTTTGT	CTAACCCTAATCCCCAGCA	Dufour, 2001	Poncet, 2007
394	AM231563	(tg)9	GCCGTCTCGTATCCCTCA	GAAGCCAGAAAGTCAGTCACATAG	Dufour, 2001	Poncet, 2007
442	AM231566	(ca)19 (ta)7	CGCAAATCTGAGTATCCCAAC	TGGATCAACACTGCCCTTC	Dufour, 2001	Poncet, 2007
445	AM231567	(ac)10	CCACAGCTTGAATGACCAGA	AATTGACCAAGTAATCACCGACT	Dufour, 2001	Poncet, 2007
455	AM408676	(ac)9 (at)9 (tttc)4	TGAAAGATGACTTTTGACTTGCTT	TCGCAATATCCTTGCTTGTCT	Dufour, 2001	Poncet, 2007
456	AM231568	(ac)14	TGGTTGTTTTCTTCCATCAATC	TCCAGTTTCCCACGCTCT	Dufour, 2001	Poncet, 2007
461	AM231570	(ac)9	CGGCTGTGACTGATGTG	AATTGCTAAGGGTCGAGAA	Dufour, 2001	Poncet, 2007
471	AM231572	(ct)12 (ca)11	TTACCTCCCGGCCAGAC	CAGGAGACCAAGACCTTAGCA	Dufour, 2001	Poncet, 2007
495	AM231575	(ac)8	CATGGATGGGAAGGCAGT	CTTGGAAAACTTGCTGAATGTG	Dufour, 2001	Poncet, 2007
501	AM231576	(tg)8	CACCACCATCTAATGCACCT	CTGCACCAGCTAATTCAAGC	Dufour, 2001	Poncet, 2007

755	AJ308755	(ca)20	CCCTCCCTCTTTCTCCTCTC	TCTGGGTTTTCTGTGTTCTCG	Rovelli, 2000	Poncet, 2004
774	AJ308774	(ct)5 (ca)7	GCCACAAGTTTCGTGCTTTT	GGGTGTCGGTGTAGGTGTATG	Rovelli, 2000	Poncet, 2004
779	AJ308779	(tg)17	TCCCCCATCTTTTTCTTTCC	GGGAGTGTTTTTGTGTTGCTT	Rovelli, 2000	Poncet, 2004
784	AJ308784		TTGCTTGCTTGTTCTGTTAT	TGACACGAGAGTTAGAAATGA	Rovelli, 2000	Poncet, 2004
837	AJ308837	(tg)16 (ga)11	CTCGCTTTCACGCTCTCTCT	CGGTATGTTCCTCGTTCCTC	Rovelli, 2000	Poncet, 2004
873	AJ308873		ACACATACATACCCCAAATGC	TTTCAGCAATACCTTGTCTATCAA	Rovelli, 2000	Poncet, 2004
119559	Not available	(A)10	GCGAGCCAGATAATCTCCAA	TCCAAGGAAAGAGAGAAACG	Nestlé R&D, Tours, SOL Genomic Network	Nestlé R&D, Tours
119699	Not available	(AT)5	GCCGTGGTGGAAGATGTACT	CGAGTTCACCAAGAACGTCA	Nestlé R&D, Tours, SOL Genomic Network	Nestlé R&D, Tours
120329	Not available	confidential	ACTCTTGGCGTTGAATTGCT	GGCTCCTTGTGTTTGGGTAA	Nestlé R&D, Tours, SOL Genomic Network	Nestlé R&D, Tours
123106	Not available	(AGA)5	GAGACGTGGTTCGTGCTGTA	GTAAATCGCAGGCTAAAGCG	Nestlé R&D, Tours, SOL Genomic Network	Nestlé R&D, Tours
DL011	AJ871890	(gct)4 (cat)8	ATACATAAGCAAGCACTGA	CAGAACAAATGAAATGGA	Leroy, 2005	Leroy, 2005
DL013	AJ871892	(ca)6 (ct)8	AGAGGGATGTCAGCATAA	ATTTGTGTTTGGTAGATGTG	Leroy, 2005	Leroy, 2005
DL025	AJ871904	(c)17	TTGTTGAGAGTGGAGGA	CCAAAGACAGTGCAGTAA	Leroy, 2005	Leroy, 2005
DL026	AJ871905	(a)17	CGAGACGAGCATAAGAA	GCTGGAATGAAGAATGTAG	Leroy, 2005	Leroy, 2005
R325	Not available	(GA)23	CCTTGTTGTTGGGGAATGTC	GGCTGTTCTGGGCTTTGTG	Nestlé R&D, Tours	Nestlé R&D, Tours
SSR009	AM231580	(gaaaa)5	CAAACAAAACAGTACAATTCAATCC	ATCCCTGCGAGACCTGACTA	Geromel, 2006	Cubry, 2008
SSR010	AM231581	(att)2	CGAAAGGAACACAGGAACCA	CAGTGGTGAACTTAATCGTCCA	Geromel, 2006	Cubry, 2008

Table 3.3: Basic statistics on the 39 markers for the global sample of 293 individuals

Marker	Major Allele Frequency	Allele Number	Availability	Gene Diversity	Hobs
BES2M2	0.37	9.00	0.90	0.78	0.28
123106	0.54	8.00	0.99	0.63	0.37
DL011	0.43	11.00	1.00	0.75	0.41
DL013	0.57	11.00	0.91	0.64	0.22
DL025	0.42	6.00	1.00	0.71	0.03
DL026	0.29	11.00	0.97	0.79	0.07
R325	0.17	18.00	0.86	0.90	0.48
SSR10	0.83	4.00	1.00	0.29	0.23
SSR9	0.31	5.00	1.00	0.73	0.52
119559	0.72	4.00	0.91	0.42	0.12
119699	0.40	6.00	0.95	0.68	0.29
120329	0.18	13.00	0.90	0.87	0.53
257	0.53	10.00	0.94	0.65	0.28
309	0.48	10.00	0.86	0.68	0.21
313	0.38	6.00	0.96	0.69	0.40
328	0.44	18.00	0.99	0.77	0.30
341	0.44	14.00	0.96	0.76	0.34
355	0.32	15.00	0.99	0.84	0.60
356	0.50	14.00	0.97	0.71	0.57
364	0.37	13.00	0.91	0.80	0.46
367	0.44	14.00	0.93	0.73	0.37
368	0.21	19.00	0.99	0.89	0.55
384	0.39	10.00	0.99	0.70	0.26
392	0.24	20.00	0.99	0.86	0.54
394	0.65	9.00	0.91	0.55	0.41
442	0.15	18.00	0.97	0.91	0.55
445	0.48	9.00	0.98	0.66	0.20
455	0.32	16.00	0.94	0.82	0.57
456	0.13	19.00	0.96	0.91	0.56
461	0.52	13.00	0.90	0.70	0.27
471	0.36	18.00	0.95	0.80	0.32
495	0.61	9.00	0.85	0.56	0.19
501	0.21	18.00	0.98	0.87	0.33
755	0.36	14.00	0.97	0.79	0.33
774	0.45	7.00	0.97	0.70	0.34
779	0.30	13.00	0.98	0.84	0.53
784	0.38	10.00	0.96	0.71	0.44
837	0.28	16.00	0.95	0.84	0.35
873	0.88	4.00	0.97	0.22	0.18
Mean*	0.41	11.85	0.95	0.72	0.36

***** Raw values not tested by bootstrapping. Hobs stand for Observed Heterozygosity.

Table 3.4: Diversity indices on the three levels of structure investigation

		Number of individuals	Major Allele Frequency	Allele Number	Gene Diversity	Hobs
1st level (2 groups)						
Congolese	**Mean***	123	0.40	10.18	0.73	0.40
	SD		*0.02*	*0.63*	*0.02*	*0.03*
	2.5% l.b.		*0.36*	*8.87*	*0.69*	*0.36*
	97.5% u.b.		*0.45*	*11.36*	*0.76*	*0.45*
Guinean	**Mean***	158	0.59	7.40	0.53	0.30
	SD		*0.04*	*0.59*	*0.04*	*0.03*
	2.5% l.b.		*0.52*	*6.26*	*0.44*	*0.24*
	97.5% u.b.		*0.66*	*8.54*	*0.60*	*0.37*
2nd level (4 groups)						
Other Congolese	**Mean***	78	0.48	8.96	0.65	0.41
	SD		*0.03*	*0.63*	*0.03*	*0.03*
	2.5% l.b.		*0.42*	*7.74*	*0.60*	*0.35*
	97.5% u.b.		*0.54*	*10.23*	*0.71*	*0.47*
Nana	**Mean***	43	0.62	4.74	0.49	0.38
	SD		*0.03*	*0.27*	*0.03*	*0.04*
	2.5% l.b.		*0.57*	*4.21*	*0.43*	*0.31*
	97.5% u.b.		*0.68*	*5.28*	*0.55*	*0.45*
Other Guinean	**Mean***	122	0.62	6.90	0.50	0.30
	SD		*0.04*	*0.60*	*0.04*	*0.03*
	2.5% l.b.		*0.55*	*5.79*	*0.42*	*0.24*
	97.5% u.b.		*0.69*	*8.18*	*0.57*	*0.38*
Pelezi	**Mean***	34	0.68	3.08	0.39	0.31
	SD		*0.03*	*0.22*	*0.04*	*0.04*
	2.5% l.b.		*0.62*	*2.64*	*0.32*	*0.23*
	97.5% u.b.		*0.75*	*3.51*	*0.46*	*0.39*
3rd level (6 groups plus Admixed)						
Other Guinean	**Mean***	122	0.62	6.90	0.50	0.30
	SD		*0.04*	*0.60*	*0.04*	*0.03*
	2.5% l.b.		*0.55*	*5.79*	*0.42*	*0.24*
	97.5% u.b.		*0.69*	*8.18*	*0.57*	*0.38*
Pelezi	**Mean***	34	0.68	3.08	0.39	0.31

	SD		*0.03*	*0.22*	*0.04*	*0.04*
	2.5% l.b.		*0.62*	*2.64*	*0.32*	*0.23*
	97.5% u.b.		*0.75*	*3.51*	*0.46*	*0.39*
Nana	**Mean***	43	0.62	4.74	0.49	0.38
	SD		*0.03*	*0.27*	*0.03*	*0.04*
	2.5% l.b.		*0.57*	*4.21*	*0.43*	*0.31*
	97.5% u.b.		*0.68*	*5.28*	*0.55*	*0.45*
B	**Mean***	31	0.67	3.79	0.44	0.38
	SD		*0.03*	*0.30*	*0.04*	*0.04*
	2.5% l.b.		*0.60*	*3.23*	*0.36*	*0.29*
	97.5% u.b.		*0.73*	*4.36*	*0.51*	*0.46*
SG1	**Mean***	8	0.61	2.92	0.47	0.42
	SD		*0.03*	*0.20*	*0.04*	*0.05*
	2.5% l.b.		*0.54*	*2.56*	*0.40*	*0.33*
	97.5% u.b.		*0.67*	*3.31*	*0.54*	*0.51*
SG2	**Mean***	34	0.52	7.41	0.62	0.44
	SD		*0.03*	*0.55*	*0.04*	*0.04*
	2.5% l.b.		*0.46*	*6.33*	*0.55*	*0.37*
	97.5% u.b.		*0.59*	*8.54*	*0.68*	*0.51*
Admixed	**Mean***	21	0.41	8.18	0.73	0.55
	SD		*0.02*	*0.52*	*0.02*	*0.03*
	2.5% l.b.		*0.37*	*7.21*	*0.69*	*0.49*
	97.5% u.b.		*0.46*	*9.26*	*0.77*	*0.61*

* Values tested by 1000 reps of a bootstrap procedure except for He (no test done), SD stands for standard deviation over the 1000 reps. 2.5% l.b. and 97.5% u.b. are respectively the lower and upper boundaries of the 95% confidence interval calculated over 1000 bootstraps. Hobs stands for Observed Heterozygosity

Table 3.5: Diversity indices for the Other Guinean group using population information and summary for 6 groups using population information for Other Guinean

		Number of individuals	Major Allele Frequency	Allele Number	Gene Diversity	Hobs
Other Guinean (using population information)						
Fourougbankoro	**Mean***	21	0.65	4.01	0.45	0.32
	SD		*0.04*	*0.35*	*0.04*	*0.04*
	2.5% l.b.		*0.59*	*3.36*	*0.37*	*0.24*
	97.5% u.b.		*0.72*	*4.69*	*0.54*	*0.40*
Cultivated Ivorian	**Mean***	16	0.61	4.71	0.50	0.34
	SD		*0.04*	*0.43*	*0.04*	*0.04*
	2.5% l.b.		*0.54*	*3.90*	*0.41*	*0.26*
	97.5% u.b.		*0.68*	*5.56*	*0.58*	*0.41*
Ira1	**Mean***	8	0.71	2.75	0.38	0.31
	SD		*0.03*	*0.19*	*0.04*	*0.05*
	2.5% l.b.		*0.65*	*2.38*	*0.31*	*0.22*
	97.5% u.b.		*0.77*	*3.13*	*0.46*	*0.39*
Ira2	**Mean***	17	0.70	3.33	0.39	0.31
	SD		*0.03*	*0.28*	*0.04*	*0.05*
	2.5% l.b.		*0.63*	*2.79*	*0.31*	*0.22*
	97.5% u.b.		*0.76*	*3.90*	*0.47*	*0.41*
Pine	**Mean***	27	0.65	3.98	0.45	0.27
	SD		*0.03*	*0.36*	*0.04*	*0.03*
	2.5% l.b.		*0.58*	*3.33*	*0.37*	*0.20*
	97.5% u.b.		*0.72*	*4.74*	*0.53*	*0.33*
Mouniandougou	**Mean***	31	0.70	3.21	0.39	0.27
	SD		*0.04*	*0.25*	*0.04*	*0.04*
	2.5% l.b.		*0.62*	*2.74*	*0.29*	*0.19*
	97.5% u.b.		*0.76*	*3.72*	*0.47*	*0.34*
Sabregue	***No computation done (only one individual)***					
Unknown	***No computation done (only one individual)***					

Summary 6 Groups plus populations for Other Guinean							
Congolese	**Nana**	**Mean**	43	0.62	4.74	0.49	0.38
	B	**Mean**	31	0.67	3.79	0.44	0.38
	SG1	**Mean**	8	0.61	2.92	0.47	0.42
	SG2	**Mean**	34	0.52	7.41	0.62	0.44
Guinean	**Fourougbankoro**	**Mean**	21	0.65	4.01	0.45	0.32
	Cultivated Ivorian	**Mean**	16	0.61	4.71	0.50	0.34
	Ira1	**Mean**	8	0.71	2.75	0.38	0.31
	Ira2	**Mean**	17	0.70	3.33	0.39	0.31
	Pine	**Mean**	27	0.65	3.98	0.45	0.27
	Mouniandougou	**Mean**	31	0.70	3.21	0.39	0.27
	Sabregue	***No computation done (only one individual)***					
	Unknown	***No computation done (only one individual)***					
	Pelezi	**Mean**	34	0.68	3.08	0.39	0.31
Admixed	**Admixed**	**Mean**	21	0.41	8.18	0.73	0.55

* Values tested by 1000 reps of a bootstrap procedure except for He (no test done), SD stands for standard deviation over the 1000 reps. 2.5% l.b. and 97.5% u.b. are respectively the lower and upper boundaries of the 95% confidence interval calculated over 1000 bootstraps. Hobs stands for Observed Heterozigosity.

Table 3.6: Pairwise R_{st} for the diverse levels of structure investigation

1st level (2 groups)										
	Guinean									
Congolese	0.401									
2nd level (4 groups)										
	Nana	Other Congolese	Pelezi							
Other Congolese	0.314									
Pelezi	0.647	0.455								
Other Guinean	0.494	0.461	0.259							
3rd level (6 groups)										
	SG2	Libenge	Nana	Niaouli	Pelezi					
Libenge	0.271									
Nana	0.421	0.476								
Niaouli	0.378	0.448	0.601							
Pelezi	0.529	0.632	0.643	0.626						
Other Guinean	0.499	0.590	0.495	0.511	0.259					
6 groups with population for Other Guinean										
	SG2	B	Nana	SG1	Mouniandougou	Fourougbankoro	Ira2	Ira1	Pine	Pelezi
B	0.271									
Nana	0.421	0.476								
SG1	0.378	0.448	0.601							
Mouniandougou	0.490	0.594	0.547	0.545						
Fourougbankoro	0.452	0.581	0.567	0.533	0.145					
Ira2	0.534	0.666	0.660	0.632	0.275	0.194				
Ira1	0.426	0.584	0.586	0.543	0.247	0.153	0.264			
Pine	0.504	0.598	0.563	0.521	0.043	0.112	0.208	0.211		
Pelezi	0.529	0.632	0.643	0.626	0.392	0.250	0.478	0.211	0.336	
Cultivated Ivorian	0.459	0.575	0.542	0.511	0.110	0.080	0.126	0.040^{ns}	0.053	0.279

^{ns} = not significant at a 5% threshold

Table 3.7: R_{st}-based AMOVAs and derived *F-statistics* on different levels of structure investigation

		Source of variation (Percentage)				Related *F-statistics*				
Level	AMOVA Design	Among groups	Among populations or among populations within groups	Among individuals within populations	Within individuals	*Ris*	*Rit*	*Rst*	*Rsc*	*Rct*
1st level (2 clusters)	2 populations (Guinean and Congolese) corresponding to the 1st level of structure	*NA*	40.1	18.29	41.62	0.30529	0.58383	0.40095	*NA*	*NA*
2nd level (4 clusters)	2 groups (Guinean and Congolese), 2 populations per group: Pelezi and Other Guinean for the Guinean group, Nana and Other Congolese for the Congolese group	27.13 ^{ns}	21.54	10.1	41.23	0.19682	0.58775	*NA*	0.29558	0.27135^{ns}
3rd level (6 clusters)	2 groups (Guinean and Congolese), 2 populations for Guinean: Pelezi and Other Guinean, 4 populations for Congolese: Nana, Libenge and SG2	27.38 ^{ns}	26.47	5.4	40.75	0.11692	0.59246	*NA*	0.36449	0.27381^{ns}

Other Guinean (6 clusters)	1 group (Other Guinean), 6 populations	*NA*	14.03	5.93	80.04	0.06897	0.19958	0.14028	*NA*	*NA*
3rd level with population information for Other Guinean (11 clusters)	2 groups (Congolese and Guinean), 4 populations for Congolese and 7 populations for Guinean	35.58	20.36	2.86	41.2	0.06498	0.58801	NA	0.31605	0.35577

^{ns} : not significant values at 5% threshold

Table 3.8: R_{is} per population on different levels of structure investigation

	Cluster	R_{is}
1st level (2 clusters)	Congolese	0.355
	Guinean	0.249
2nd level (4 clusters)	Nana	0.070(ns)
	Other Congolese	0.256
	Pelezi	0.087(ns)
	Other Guinean	0.178
3rd level (6 clusters)	Nana	0.070(ns)
	Libenge	0.011(ns)
	Niaouli	0.144(ns)
	SG2	0.070(ns)
	Pelezi	0.087(ns)
	Other Guinean	0.178
Other Guinean (6 clusters)	Pine	0.056
	Cultivated Ivorian	0.146
	Ira1	0.003
	Ira2	0.000
	Fourougbankoro	0.105
	Mouniandougou	0.111
3rd level with population information for Other Guinean (11 clusters)	SG2	0.070(ns)
	B	0.011(ns)
	Nana	0.070(ns)
	SG1	0.144(ns)
	Mouniandougou	0.111
	Fourougbankoro	0.105(ns)
	Ira2	0.000(ns)
	Ira1	0.008(ns)
	Pine	0.056(ns)
	Pelezi	0.087(ns)
	Cultivated Guinean	0.146

[ns]: not significant values at 5% threshold

Données supplémentaires : voir Annexes A.3.1 à A.3.4

Analyse de diversité de génotypes d'une population d'amélioration

Introduction

Ce travail a pour but de situer la diversité de génotypes d'origine « conilon » dans la diversité globale de *C. canephora* : groupes Guinéens (G et Pélézi), et groupes Congolais (SG1, SG2, B, C et UW), afin d'évaluer notre capacité à caractériser de manière efficace les collections et la diversité présente en leur sein. Ce travail a été réalisé en collaboration avec le centre de recherche Nestlé de Tours. Le nom Conilon serait *a priori* dérivé du nom Kouilou ou quillou (de Madagascar) qui désignait à l'origine des génotypes cultivés en Côte d'Ivoire originaires de la façade atlantique du Gabon ou du Congo Brazzaville, ayant transité par Java (Cramer, 1957). Cette donnée nous amène à penser les retrouver proches génétiquement de la population Niaouli, identifiée comme appartenant au groupe SG1.

Matériel et méthodes

Matériel végétal

22 individus d'une population Conilon utilisée par Nestlé et 9 témoins de diversité fournis par le CIRAD ont été génotypés à l'aide de 16 marqueurs microsatellites au centre de recherche Nestlé de Tours. L'ensemble de ces génotypes est listé dans le tableau 3.9.

Tableau 3.9 : liste des génotypes Conilon inclus dans l'étude et témoins utilisés

Code	**Groupe**	**Origine**
C3005	Congolais SG1	CIRAD
1648-1	Congolais SG1	CIRAD
C2011	Congolais SG2	CIRAD
C2014	Congolais SG2	CIRAD
C5003	Congolais SG2	CIRAD
C4047	Congolais C	CIRAD
g1003	Guinéens sauvage	CIRAD
g7001	Guinéens sauvage	CIRAD
c1017	Congolais B	CIRAD
116M	Conilon	Nestlé
100M	Conilon	Nestlé
120M	Conilon	Nestlé
O3P	Conilon	Nestlé
154P	Conilon	Nestlé

106T	Conilon	Nestlé
07M	Conilon	Nestlé
128M	Conilon	Nestlé
112M	Conilon	Nestlé
Confidentiel	Conilon	Confidentiel
Confidentiel	Conilon	Confidentiel
Confidentiel	Conilon	Confidentiel
Confidentiel	Conilon	Confidentiel
Confidentiel	Conilon	Confidentiel
Confidentiel	Conilon	Confidentiel
Confidentiel	Conilon	Confidentiel
Confidentiel	Conilon	Confidentiel
Confidentiel	Conilon	Confidentiel
Confidentiel	Conilon	Confidentiel
Confidentiel	Conilon	Confidentiel
Confidentiel	Conilon	Confidentiel
Confidentiel	Conilon	Confidentiel

Une partie de ces génotypes nous est totalement inconnue pour des raisons de confidentialité, nous n'avons pour ceux-ci que l'information de la population d'origine, la population Conilon.

Cette étude nous a permis d'incorporer les génotypes Conilon dans 2 études de diversité à plus large échelle. Une première regroupe tous les génotypes des études de diversité effectuées au CIRAD (523 génotypes au total, mais avec un nombre de marqueurs restreint) et une seconde concerne les groupes de diversité SG1, SG2, C et Guinéens (373 individus).

Marqueurs utilisés

Les 16 marqueurs microsatellites génotypés sur cette population ont été identifiés et fournis par le CIRAD. La liste de ces marqueurs est donnée en tableau 3.10. 10 de ces marqueurs sont communs avec l'étude de diversité présentée dans le manuscrit en première partie de ce chapitre, les 6 autres ont été utilisés pour des génotypages complémentaires sur l'ensemble des génotypes de cette précédente étude à l'exception de la population Libengé, mais en y incorporant des génotypes sauvages originaires d'Ouganda. Le choix des marqueurs a principalement été basé sur la taille du motif et la qualité d'amplification.

Tableau 3.10 : Liste des 16 marqueurs utilisés dans cette étude

Nom	**Motif microsatellite (nombre de répétitions dans clone 126)**	**Présent dans l'étude précédente**
257	(CA) 9	Oui
337	(TG) 8	Non
338 (=368)	(TG) 3 (TT) 1 (TG) 13	Oui
358	(CA) 11	Non
360	(CA) 10	Non
442	(CA) 19	Oui
445	(AC) 10	Oui
461	(AC) 9	Oui
477	(AC) 16	Non
501	(TG) 8	Oui
753	(CA) 15	Non
779	(TG) 17	Oui
782	(GT) 15	Non
837	(TG) 16 (TA) 11	Oui
DL011	(GCT) 4 (CAT) 8	Oui
DL013	(CA) 6 (CT) 8	Oui

Fusion des données

Nous avons pu fusionner 15 de ces marqueurs par un simple recalage des longueurs d'allèles. En revanche une incertitude trop importante sur les résultats obtenus sur le marqueur 442 ne nous a pas permis d'incorporer celui-ci à notre analyse. Le tableau 3.11 donne la relation entre les tailles relatives des jeux de données dans le sens Nestlé -> CIRAD.

Tableau 3.11 : Différence entre les tailles relatives obtenues au CIRAD et à Nestlé

Nom	**Différence**
782	+18
837	+19
DL011	+20
DL013	+20
257	+24
337	+18
338	+16
358	+18
360	+17
442	Non Utilisé
445	+22
461	+24
477	+20
501	+23
753	+22
779	+26

Ces différences de taille d'allèles se sont révélées pour la plupart systématique et ont été corrigées à l'aide d'un tableur. La validité des corrections a été vérifiée à l'aide des témoins présents dans les différentes études.

Nous avions à disposition un certain nombre d'étude de diversité au CIRAD, dont principalement celle exposé en première partie de ce chapitre ainsi que celle portant sur des génotypes collectés en Ouganda et donnée en Annexe A.3.5 (Musoli et al, 2008, soumis à *Genome*).

Nous avons pu utiliser 7 marqueurs sur l'ensemble des génotypes. Par ailleurs 15 marqueurs sont utilisés sur les génotypes des populations C, SG2, SG1 et G que nous avions caractérisées dans l'étude précédente en y joignant, en plus de la population Conilon, des génotypes d'une population collectée à Luki en République Démocratique du Congo, descendances libres de génotypes spontanés de cette région (Bieysse, communication personnelle).

Analyse de diversité

Une première analyse graphique a été réalisée à l'aide du logiciel DARwin 5.0 développé au CIRAD (Perrier *et al*, 2003; Perrier & Jacquemoud-Collet, 2006) qui permet de réaliser une analyse en composante principale basée sur un tableau de dissimilarités entre génotypes calculé à l'aide d'un index simple-matching (analogue au nombre d'allèles partagés). Ce même tableau peut ensuite être utilisé comme base au calcul d'un arbre selon divers algorithmes, dont celui de Neighbour-Joining.

Nous avons par ailleurs calculé quelques statistiques descriptives à l'aide du logiciel PowerMarker 3.25 (Liu & Muse, 2005).

Résultats

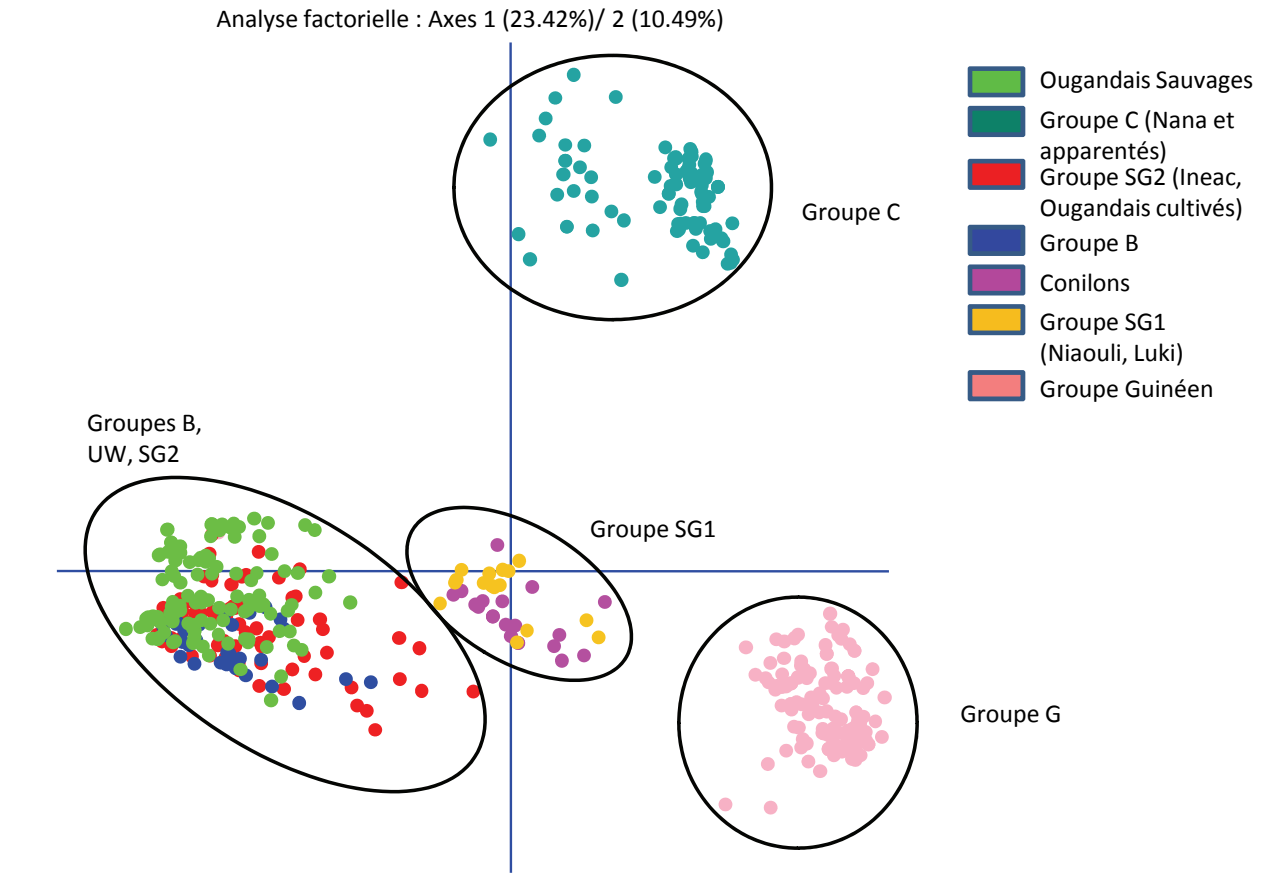
Analyse factorielle et représentation NJ

519 individus et 7 marqueurs

Une première analyse incorporant l'ensemble des génotypes sur les axes 1 et 2 et les axes 1 et 3 est donnée en figure 3.6. L'arbre NJ non raciné dérivé est donné en figure 3.7.

L'analyse de ces deux figures indique que les Conilon sont apparentés au groupe SG1 et *a priori* uniquement à celui-ci. L'utilisation de seulement 7 marqueurs suffit à discriminer

aisément les groupes de diversité SG1, C et Guinéens. Seuls les groupes SG2, B et UW (génotypes de RDC et sauvages d'Ouganda) sont plus difficilement discernables et semblent former un seul grand groupe sur l'analyse factorielle. Ceci peut être expliqué par la proximité génétique de ces groupes ainsi que par le poids occupé par les autres groupes dans l'analyse factorielle. Sur l'arbre en revanche une séparation commence à se dessiner, confirmant l'originalité des populations sauvages d'Ouganda. Ces résultats sont en concordances avec l'hypothèse précédemment énoncée d'une proximité importante entre les groupes SG2 et B. Il semble cependant qu'un nombre plus important de marqueurs serait utile pour préciser les relations génétiques existantes entre ces différents groupes.



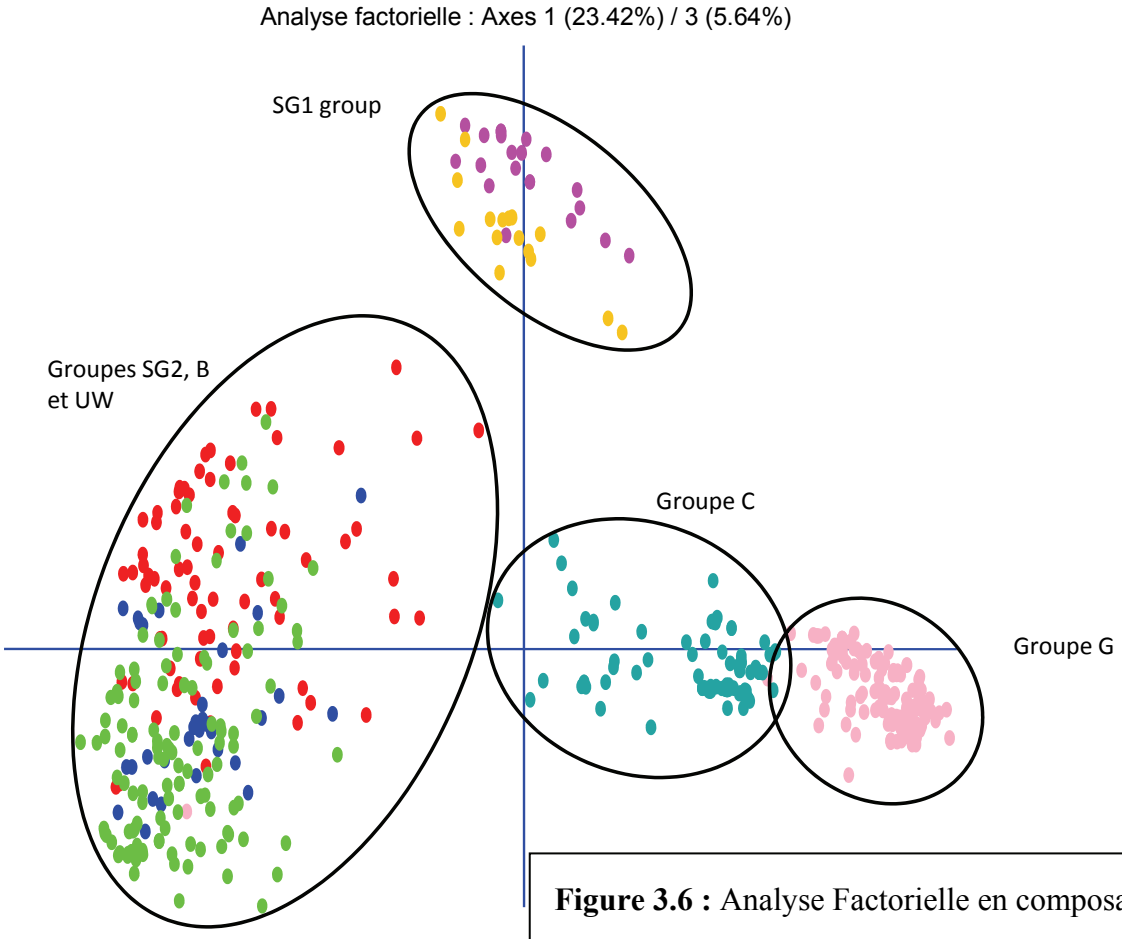


Figure 3.6 : Analyse Factorielle en composante principale sur Tableau de Dissimilarités (519 individus, 7 marqueurs)

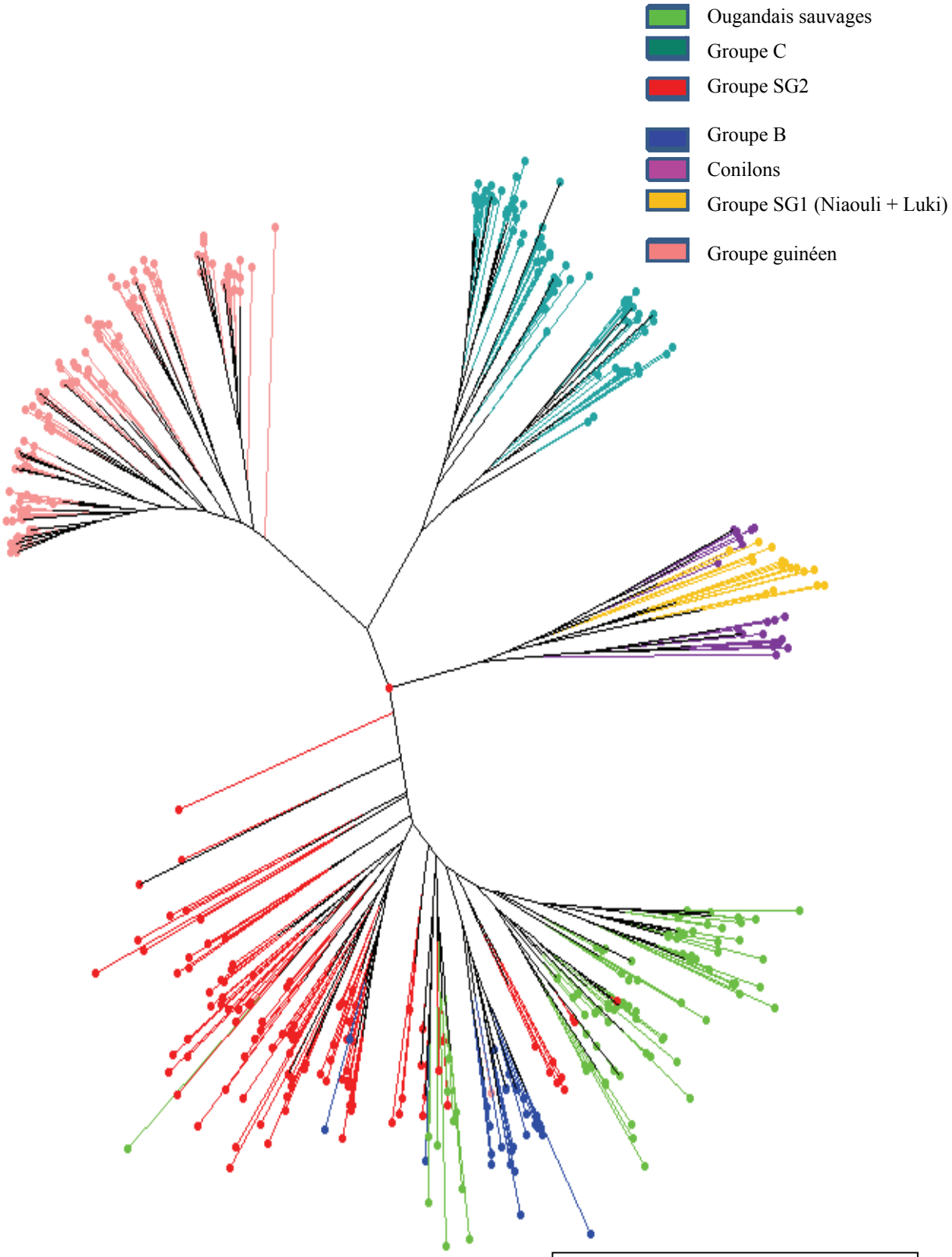


Figure 3.7 : Arbre Neighbour-Joining basé sur Tableau de Dissimilarités (519 individus, 7 marqueurs)

373 individus et 15 marqueurs

Les représentations graphiques de l'analyse factorielle avec l'ensemble des groupes de diversité étudiés (SG1, SG2, C et Guinéens) sont donnés en figure 3.8. Une analyse portant uniquement sur les groupes Congolais (SG1, SG2 et C) est donnée en figure 3.9. Enfin un arbre NJ comprenant les 373 individus est donné en figure 3.10.

Ces analyses confirment les résultats obtenus plus haut. Les Conilon font partie intégrante du groupe SG1 et proviennent vraisemblablement d'une même zone géographique située au niveau de la façade atlantique du Gabon, du Congo-Brazzaville ou de RDC (vraisemblablement le bassin du Niaouli pour l'origine togolaise Niaouli).

Au vu de ces résultats, la diversité des Conilon semble restreinte par rapport à l'ensemble de l'espèce, ne constituant *a priori* qu'une partie de l'origine SG1.

Un arbre NJ concernant uniquement le groupe SG1 est donné en figure 3.11.On peut se rendre compte sur celui-ci que 15 marqueurs microsatellites arrivent à discerner les 3 populations différentes (Niaouli, Luki et Conilon) sans aucune ambiguité. Ce résultat semble indiquer que ces 15 marqueurs sont assez informatifs pour permettre une étude globale de la diversité pouvant permettre d'aller jusqu'à la distinction de populations différentes au sein d'une même origine géographique.

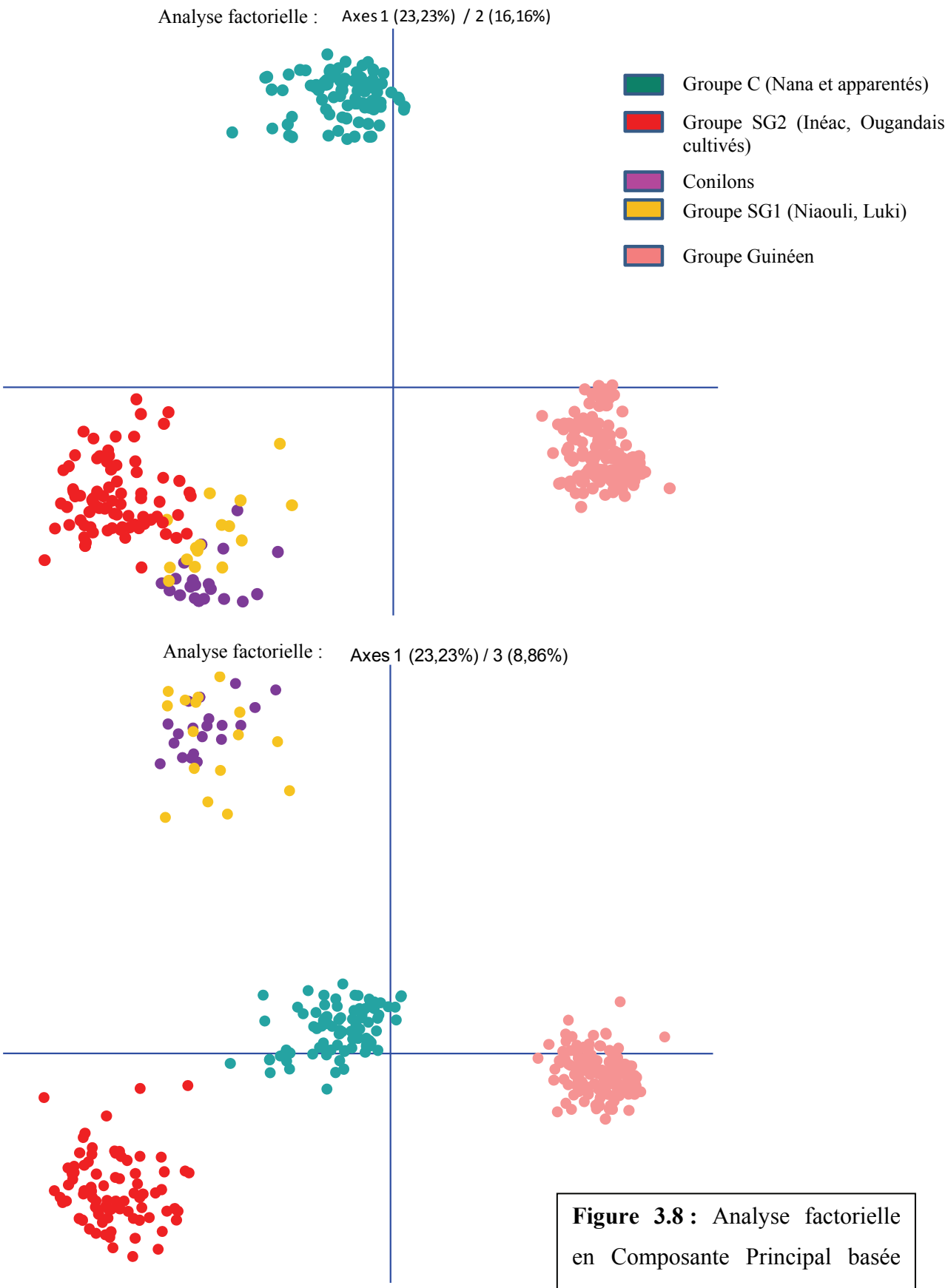


Figure 3.8 : Analyse factorielle en Composante Principal basée sur Tableau de Dissimilarités (373 individus et 15 marqueurs)

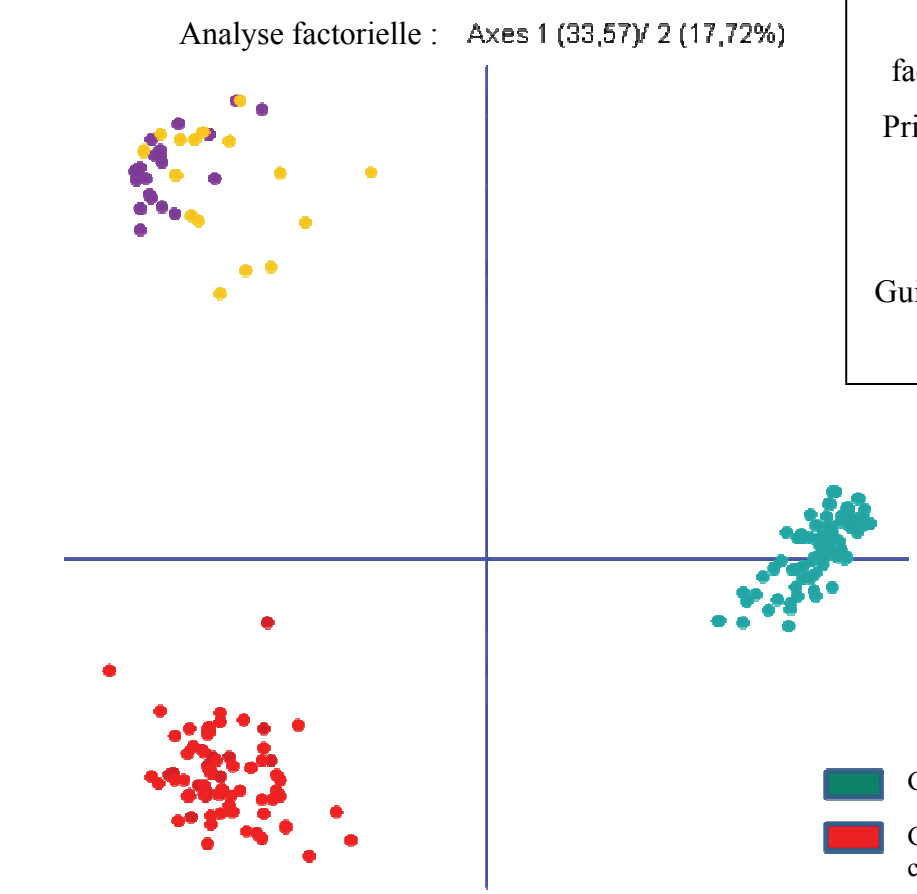


Figure 3.9 : Analyse factorielle en Composante Principal basée sur Tableau de Dissimilarités (373 individus moins les Guinéens soit 199 génotypes pour 15 marqueurs)

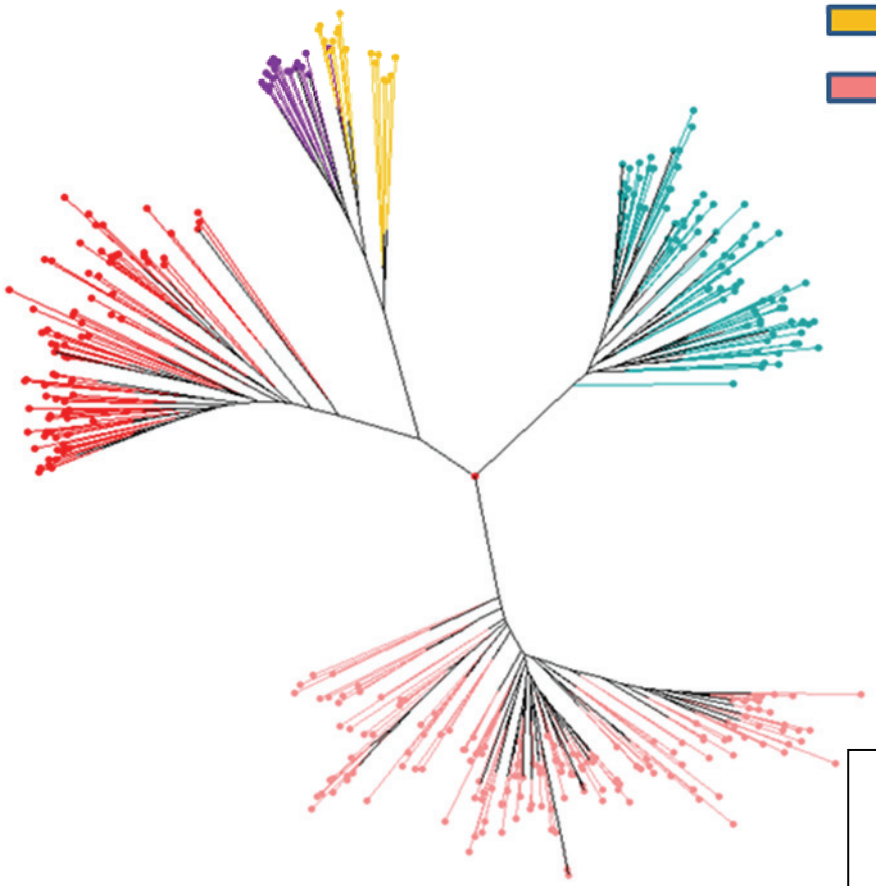


Figure 3.10 : Arbre Neighbour-Joining basé sur Tableau de Dissimilarités (373 individus pour 15 marqueurs)

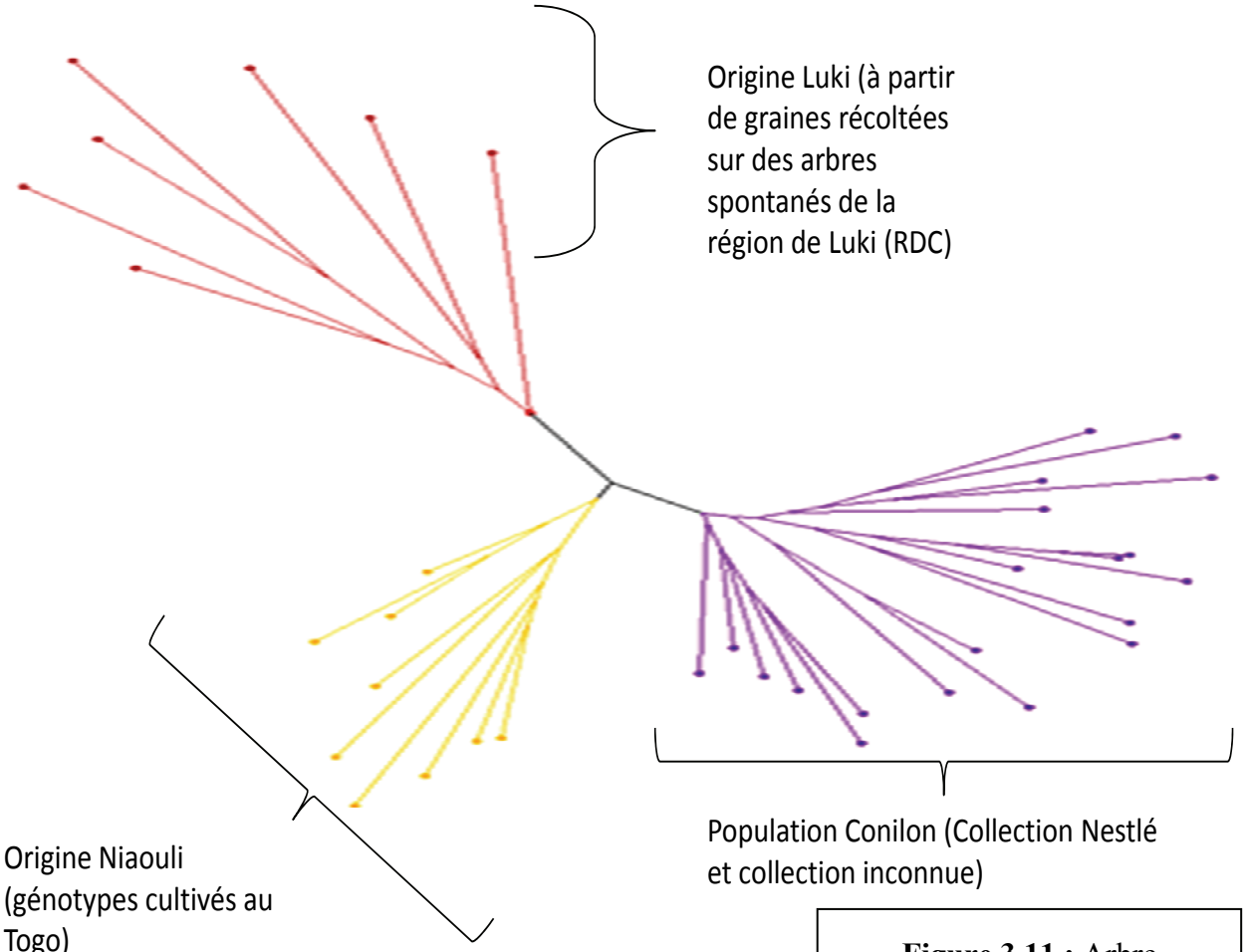


Figure 3.11 : Arbre Neighbour-Joining basé sur Tableau de Dissimilarités (38 individus SG1 pour 15 marqueurs)

Statistiques descriptives

Sur les marqueurs

Nous avons réalisé un calcul de statistiques descriptives sur 7 et 15 marqueurs respectivement. Celles-ci comprennent le nombre total d'allèles, l'hétérozygotie observée (Ho) et deux analogues de l'hétérozygotie attendue, la « Gene Diversity » (Gene Diversity) et le « Polymorphism Information Content » (PIC). Des moyennes et intervalles de confiance à 95% ont été calculés à l'aide de 1000 bootstraps pour l'ensemble des valeurs, ces tests sont donnés en italiques sur les tableaux suivants. Les statistiques calculées pour les 7 marqueurs et les 519 individus sont données en Tableau 3.12, celles calculées pour les 15 marqueurs et 373 individus en Tableau 3.13.

Tableau 3.12 : statistiques descriptives pour 7 marqueurs et 519 individus. *2.5% b.inf. et 97.5% b.sup.* = bornes de l'intervalle de confiance à 95%

Marqueur	Nombre d'Allèles	Pourcentage d'amplification	Gene Diversity	Hétérozygotie observées (Ho)	PIC
DL011	10	0.943	0.760	0.319	0.726
DL013	15	0.955	0.788	0.290	0.759
338	25	0.996	0.908	0.627	0.900
445	12	0.968	0.633	0.194	0.568
461	23	0.981	0.880	0.486	0.871
501	21	0.987	0.829	0.369	0.815
837	19	0.975	0.879	0.447	0.868
Moyenne	17.857	0.972	0.811	0.390	0.787
Moyenne (1000 bootstraps)	*17.866*	*0.972*	*0.809*	*0.393*	*0.786*
Ecart type (1000 bootstraps)	*1.985*	*0.006*	*0.034*	*0.050*	*0.039*
2.5% b.inf.	*13.857*	*0.960*	*0.740*	*0.294*	*0.708*
97.5% b.sup.	*21.714*	*0.984*	*0.871*	*0.493*	*0.855*

Tableau 3.13 : statistiques descriptives pour 15 marqueurs et 373 individus. *2.5% b.inf. et 97.5% b.sup.* = bornes de l'intervalle de confiance à 95%

Marqueur	Nombre d'Allèles	Pourcentage d'amplification	Gene Diversity	Hétérozygotie observées (Ho)	PIC
DL011	7	0.992	0.780	0.354	0.750
DL013	13	0.984	0.745	0.276	0.721
257	10	0.984	0.619	0.241	0.564
337	11	0.995	0.696	0.125	0.648
338	21	0.995	0.889	0.607	0.879
358	15	0.992	0.827	0.271	0.806
360	13	0.982	0.881	0.470	0.869
445	10	0.953	0.529	0.169	0.488
461	21	0.960	0.814	0.486	0.799

477	16	0.989	0.851	0.557	0.835
501	17	0.992	0.890	0.455	0.880
753	15	0.982	0.835	0.602	0.820
779	15	0.982	0.863	0.559	0.848
782	6	0.987	0.234	0.045	0.212
837	16	0.976	0.840	0.414	0.824
Moyenne	13.733	0.983	0.753	0.376	0.729
Moyenne	*13.746*	*0.983*	*0.748*	*0.377*	*0.730*
Ecart Type	*1.097*	*0.003*	*0.049*	*0.045*	*0.046*
2.5% b. inf.	*11.667*	*0.977*	*0.641*	*0.291*	*0.635*
97.5% b.sup.	*15.867*	*0.988*	*0.828*	*0.460*	*0.810*

Pour les 2 échantillons la diversité est importante avec respectivement 17,9 et 13,7 allèles en moyenne et une Gene Diversity de 0,811 et 0,753. Le premier échantillon semble contenir une diversité plus importante ce qui est conforme avec le fait qu'il contienne un plus grand nombre de groupes. Dans l'ensemble, les marqueurs choisis semblent informatifs sur la diversité avec un nombre d'allèles minimum de 6 et un maximum de 25. La perte de 2 à 4 allèles sur les 7 marqueurs du premier échantillon vers le second est cohérente avec la perte de 2 groupes de diversité (B et UW) qui sont proches des SG2 et n'apportent que peu d'allèles spécifiques.

L'amplification sur l'ensemble a été très correcte avec plus de 97% et 98% de données présentes pour le premier et le second échantillon respectivement.

Sur les populations ou groupes de diversité

Les statistiques descriptives moyennes calculées par groupe de diversité et/ou par population sont données en Tableau 3.14 pour les 519 individus et en Tableau 3.15 pour les 373 individus.

Tableau 3.14 : statistiques descriptives par population pour 7 marqueurs. *2.5% b.inf. et 97.5% b.sup.* = bornes de l'intervalle de confiance à 95%.

Pop		Nombre d'Allèles	Pourcentage d'amplification	Gene Diversity (analogue He)	Hétérozygotie Observée	PIC
B	**Moyenne**	**4.857**	**0.952**	**0.439**	**0.358**	**0.398**
	Moyenne	*4.866*	*0.952*	*0.434*	*0.360*	*0.393*
	Ecart Type	*0.821*	*0.034*	*0.089*	*0.090*	*0.079*
	2.5% b.inf.	*3.286*	*0.874*	*0.247*	*0.185*	*0.232*
	97.5% b.sup.	*6.429*	*1.000*	*0.595*	*0.538*	*0.538*
C	**Moyenne**	**5.000**	**0.978**	**0.490**	**0.363**	**0.434**
	Moyenne	*5.030*	*0.977*	*0.488*	*0.361*	*0.433*
	Ecart Type	*0.938*	*0.008*	*0.070*	*0.067*	*0.068*
	2.5% b.inf.	*3.571*	*0.961*	*0.349*	*0.223*	*0.305*
	97.5% b.sup.	*7.143*	*0.992*	*0.620*	*0.489*	*0.571*

G	**Moyenne**	**6.000**	**0.993**	**0.440**	**0.281**	**0.419**
	Moyenne	*6.046*	*0.993*	*0.434*	*0.282*	*0.417*
	Ecart Type	*1.232*	*0.003*	*0.116*	*0.069*	*0.112*
	2.5% b.inf.	*3.857*	*0.988*	*0.206*	*0.151*	*0.201*
	97.5% b.sup.	*8.571*	*0.998*	*0.655*	*0.423*	*0.662*
SG1 tous	**Moyenne**	**5.857**	**0.981**	**0.657**	**0.541**	**0.606**
	Moyenne	*5.882*	*0.981*	*0.654*	*0.542*	*0.606*
	Ecart Type	*1.048*	*0.010*	*0.046*	*0.080*	*0.048*
	2.5% b.inf.	*4.286*	*0.959*	*0.567*	*0.387*	*0.513*
	97.5% b.sup.	*8.143*	*0.996*	*0.741*	*0.694*	*0.701*
Conilons	**Moyenne**	**3.714**	**1.000**	**0.569**	**0.539**	**0.513**
	Moyenne	*3.736*	*1.000*	*0.566*	*0.541*	*0.514*
	Ecart Type	*0.382*	*0.000*	*0.086*	*0.109*	*0.078*
	2.5% b.inf.	*3.000*	*1.000*	*0.383*	*0.318*	*0.335*
	97.5% b.sup.	*4.429*	*1.000*	*0.694*	*0.747*	*0.633*
SG1 autres (Luki et Niaouli)	**Moyenne**	**4.857**	**0.955**	**0.639**	**0.550**	**0.581**
	Moyenne	*4.869*	*0.955*	*0.638*	*0.552*	*0.579*
	Ecart Type	*0.727*	*0.024*	*0.036*	*0.063*	*0.045*
	2.5% b.inf.	*3.714*	*0.902*	*0.575*	*0.433*	*0.499*
	97.5% b.sup.	*6.429*	*0.991*	*0.714*	*0.677*	*0.674*
Luki	**Moyenne**	**3.571**	**1.000**	**0.583**	**0.571**	**0.508**
	Moyenne	*3.582*	*1.000*	*0.583*	*0.571*	*0.507*
	Ecart Type	*0.398*	*0.000*	*0.026*	*0.056*	*0.032*
	2.5% b.inf.	*2.857*	*1.000*	*0.536*	*0.469*	*0.449*
	97.5% b.sup.	*4.286*	*1.000*	*0.634*	*0.673*	*0.571*
Niaouli	**Moyenne**	**3.000**	**0.921**	**0.557**	**0.532**	**0.478**
	Moyenne	*3.008*	*0.919*	*0.557*	*0.530*	*0.478*
	Ecart Type	*0.202*	*0.045*	*0.016*	*0.071*	*0.023*
	2.5% b.inf.	*2.571*	*0.825*	*0.527*	*0.401*	*0.432*
	97.5% b.sup.	*3.429*	*1.000*	*0.587*	*0.663*	*0.520*
SG2	**Moyenne**	**10.429**	**0.978**	**0.605**	**0.569**	**0.582**
	Moyenne	*10.458*	*0.979*	*0.603*	*0.569*	*0.578*
	Ecart Type	*1.480*	*0.008*	*0.092*	*0.085*	*0.089*
	2.5% b.inf.	*7.857*	*0.962*	*0.421*	*0.381*	*0.403*
	97.5% b.sup.	*13.429*	*0.993*	*0.773*	*0.720*	*0.741*
UW	**Moyenne**	**6.143**	**0.940**	**0.453**	**0.331**	**0.420**
	Moyenne	*6.156*	*0.941*	*0.459*	*0.333*	*0.415*
	Ecart Type	*1.379*	*0.028*	*0.116*	*0.088*	*0.108*
	2.5% b.inf.	*3.286*	*0.878*	*0.213*	*0.154*	*0.200*
	97.5% b.sup.	*8.857*	*0.983*	*0.666*	*0.496*	*0.607*

Tableau 3.15 : statistiques descriptives par population pour 15 marqueurs. *2.5% b.inf. et 97.5% b.sup.* = bornes de l'intervalle de confiance à 95%

Pop		Nombre d'Allèles	Pourcentage d'amplification	Gene Diversity	Hétérozygotie Observée	PIC
SG1	**Moyenne**	**5.000**	**0.984**	**0.533**	**0.455**	**0.485**
	Moyenne	*5.011*	*0.984*	*0.533*	*0.453*	*0.488*
	Ecart Type	*0.595*	*0.008*	*0.055*	*0.055*	*0.050*
	2.5% b.inf.	*3.867*	*0.967*	*0.423*	*0.346*	*0.390*
	97.5% b.sup.	*6.267*	*0.998*	*0.637*	*0.556*	*0.584*
Conilons	**Moyenne**	**3.000**	**1.000**	**0.413**	**0.418**	**0.368**
	Moyenne	*3.010*	*1.000*	*0.412*	*0.416*	*0.372*
	Ecart Type	*0.338*	*0.000*	*0.068*	*0.070*	*0.059*
	2.5% b.inf.	*2.333*	*1.000*	*0.278*	*0.285*	*0.256*
	97.5% b.sup.	*3.667*	*1.000*	*0.544*	*0.552*	*0.481*
Niaouli	**Moyenne**	**2.333**	**0.941**	**0.429**	**0.463**	**0.355**

	Moyenne	*2.334*	*0.940*	*0.430*	*0.463*	*0.356*
	Ecart Type	*0.201*	*0.030*	*0.049*	*0.065*	*0.042*
	2.5% b.inf.	*1.933*	*0.874*	*0.329*	*0.333*	*0.273*
	97.5% b.sup.	*2.733*	*0.993*	*0.517*	*0.587*	*0.435*
Luki	**Moyenne**	**3.667**	**0.990**	**0.572**	**0.554**	**0.512**
	Moyenne	*3.676*	*0.990*	*0.569*	*0.554*	*0.513*
	Ecart Type	*0.320*	*0.009*	*0.044*	*0.061*	*0.041*
	2.5% b.inf.	*3.000*	*0.971*	*0.469*	*0.421*	*0.425*
	97.5% b.sup.	*4.267*	*1.000*	*0.640*	*0.657*	*0.579*
G	**Moyenne**	**7.267**	**0.991**	**0.453**	**0.319**	**0.429**
	Moyenne	*7.302*	*0.991*	*0.449*	*0.317*	*0.431*
	Ecart Type	*0.804*	*0.002*	*0.080*	*0.058*	*0.074*
	2.5% b.inf.	*5.733*	*0.986*	*0.298*	*0.205*	*0.287*
	97.5% b.sup.	*8.933*	*0.996*	*0.604*	*0.425*	*0.570*
SG2	**Moyenne**	**7.800**	**0.980**	**0.530**	**0.442**	**0.501**
	Moyenne	*7.795*	*0.979*	*0.530*	*0.440*	*0.502*
	Ecart Type	*0.849*	*0.008*	*0.066*	*0.066*	*0.065*
	2.5% b.inf.	*6.133*	*0.961*	*0.395*	*0.310*	*0.380*
	97.5% b.sup.	*9.533*	*0.993*	*0.651*	*0.575*	*0.628*
C	**Moyenne**	**4.533**	**0.979**	**0.451**	**0.387**	**0.400**
	Moyenne	*4.552*	*0.979*	*0.450*	*0.384*	*0.400*
	Ecart Type	*0.689*	*0.006*	*0.059*	*0.059*	*0.055*
	2.5% b.inf.	*3.333*	*0.966*	*0.328*	*0.275*	*0.290*
	97.5% b.sup.	*6.067*	*0.989*	*0.568*	*0.497*	*0.508*

Les résultats obtenus sur 7 marqueurs placent le groupe SG1 comme intermédiaire en termes de nombre d'allèles alors qu'il possède la plus grande Gene Diversity. L'hétérozygotie observée est élevée et proche de celle observée chez les SG2 qui constituent le groupe le plus diverse que nous connaissons jusqu'à présent. Néanmoins le faible nombre de marqueurs et la largeur des intervalles de confiance ne permettent pas de formuler des conclusions fermes sur cette partie de l'étude.

Les valeurs calculées sur 15 marqueurs semblent plus robustes avec des écart-types inférieurs et des intervalles de confiance plus resserrés. Le groupe SG1 renferme une diversité qui semble de l'ordre de celle trouvée par exemple dans le groupe C mais qui reste inférieure à celle du groupe SG2.

Dans les 2 cas la population Conilon du groupe SG1 se place comme intermédiaire entre la population cultivée Niaouli et la population spontanée Luki. Cette population Conilon ne saurait représenter l'ensemble de la diversité contenue dans le groupe SG1.

Discussion et conclusion

Cette étude nous a permis d'étudier la place de la population Conilon dans l'ensemble de la diversité disponible de *C. canephora* au CIRAD. Nous avons ainsi montré que les

génotypes d'origine Conilon font partie intégrante d'un des groupes précédemment étudiés, les SG1, comme pouvait le laisser supposer les hypothèses sur l'origine du mot « Conilon » (Cramer, 1957). Les indications de Cramer permettent de situer l'origine de cette population au jardin botanique de Libreville, à partir duquel 33 génotypes ont été transférés à Java. Ces génotypes étaient vraisemblablement des génotypes cultivés au Gabon et au bas-Congo. La base génétique de cette population apparaît donc relativement étroite, ce qui est confirmé par l'homogénéité génétique que nous observons. De plus cette population ne correspond qu'à une origine particulière du groupe SG1. Cette hypothèse est renforcée par la diversité plus importante observée chez les caféiers spontanés de la région de Luki.

Il apparaît par ailleurs qu'aucune hybridation avec d'autres groupes de diversité comme par exemple les SG2 ou les Guinéens à proximité desquels des génotypes du groupe SG1 ont pu être cultivé n'est détectée dans nos études pour aucune des 3 populations de ce groupe. La population Conilon que nous observons n'a donc pas subi de pollution par les autres sources de diversité génétique de l'espèce.

Les résultats obtenus dans la présente étude permettent de mieux appréhender l'amélioration variétale possible de *C. canephora* ainsi que la diversité génétique disponible. Nous avons mis en évidence la place qu'occupent les génotypes Conilon dans l'ensemble de la diversité de l'espèce. Celle-ci est relativement faible et correspond à une unique population du groupe SG1, qui n'apporte que peu d'originalité par rapport aux génotypes déjà connus de celui-ci.

La cartographie génétique de *Coffea canephora* et la recherche de zones génomiques intéressantes pour l'amélioration

Introduction

Afin d'identifier des locus impliqués dans la variation de caractères phénotypiques, qu'ils soient discrets (résistances par exemple) ou continus (taille des grains, rendement etc), les approches de cartographie génétique ont prouvé leur efficacité. Dans le cadre d'un projet européen INCO (Improving Quality of African Robustas) une carte génétique intraspécifique de *C. canephora* a été initiée au CIRAD. Cette carte repose sur une descendance de 254 individus issus d'un croisement de type test-cross entre un hybride intergroupe (Congolais par Guinéen) et un génotype Guinéen. Le travail réalisé dans le cadre de cette thèse a principalement porté sur le développement et l'analyse de certains marqueurs microsatellites dans des séquences de gènes, d'extrémités de clones BAC et de séquence de clone BAC.

Etat de la carte génétique en octobre 2008

Un total de 249 marqueurs a été cartographié sur la carte génétique. Celle-ci comporte 11 groupes de liaison pour une longueur totale de 1407,6 centiMorgan. Ces marqueurs sont majoritairement des microsatellites issus de diverses sources, ainsi qu'un certain nombre de fragments de gènes candidats, notamment impliqués dans les métabolismes des sucres (Saccharose Synthétase), des lipides et des diterpènes. Une version de cette carte est donnée en Annexe A.3.6.

La recherche de QTL de qualité et de production : état des lieux

L'objectif principal du développement d'une carte génétique est la recherche de QTL, c'est-à-dire de locus impliqués dans la variation de caractères phénotypiques d'intérêt. Dans le cadre du projet ayant promu le développement de cette carte, un certain nombre d'analyses phénotypiques sur des caractères agronomiques, technologiques et sensoriels ont été réalisés. Nous avons ainsi pu mettre en évidence un certain nombre de QTL répétés sur plusieurs années et ayant des significations importantes. Ces QTL comprennent notamment des QTL de productivité, de granulométrie (taille des graines), d'acidité, d'amertume, de taux de sucres ou de caféine (Leroy et al, en préparation).

Ces régions QTL colocalisent pour certaines avec des fragments de gènes candidats que nous avons cartographiés, c'est le cas par exemple de QTL de caféine colocalisant avec un gène de Caféine Synthéthase.

Le clone BAC 111O18

Dans le cadre d'une collaboration avec le centre de recherche Nestlé de Tours, un clone BAC (Bacterial Artificial Chromosome), le clone 111O18 d'environ 180 kb, à été séquencé. Ce clone est un analogue du BAC19 de la tomate (*Solanum lycopersicum*, Crouzillat, communication personnelle). Nous avons identifié un certain nombre de marqueurs microsatellites dans cette séquence, dont certains ont été cartographiés. Nous avons ainsi pu déterminer la position de ce clone sur notre carte. Il est intéressant de noter que le Bac 111O18 colocalise avec un QTL de granulométrie. Lorsqu'on étudie le clone BAC19 de tomate celui-ci comporte un gène, *Ovate*, impliqué dans la forme du fruit (Liu *et al.*, 2002). Ce gène est un gène de régulation important pour le développement de la plante. Son homologue, situé dans ce clone BAC de caféier (résultat BLAST), semble être un bon candidat pour le QTL de granulométrie qui colocalise avec cette région.

Perspectives des études d'association et valorisation des résultats de cartographie

Ces études, et la localisation de zones génomiques intéressantes, nous permettent d'envisager des applications à court terme pour la génétique d'association. En effet de telles approches sont souvent utilisées pour affiner les intervalles de confiance des QTL voire identifier les polymorphismes responsables des variations de caractères. Néanmoins ce type d'études basées sur des gènes ou des régions candidates nécessite au préalable d'identifier ces régions.

Cette stratégie a déjà été utilisée avec succès chez de nombreuses plantes, et notamment chez la vigne où elle a permis de réduire la zone d'intérêt d'une dizaine de cM à quelques gènes (This, communication personnelle). Nous proposons donc d'utiliser les résultats obtenus en cartographie génétique et recherche de QTL pour déterminer les régions génomiques les plus susceptibles d'être étudiées en génétique d'association. Une première étude pourra être menée sur des caractères liés à la qualité ou à la productivité tels que la teneur en caféine de la graine ou la granulométrie.

Discussion – conclusion du chapitre

Nous avons étudié l'ensemble des origines géographiques disponibles de *C. canephora*, montrant l'important réservoir de diversité mobilisable pour l'amélioration dans cette espèce. Globalement cette diversité est fortement structurée avec un clivage très important entre l'ouest et le centre est de l'Afrique (Guinéens versus Congolais). Nous avons aussi mis en évidence une structuration à une échelle inférieure, notamment avec les différents groupes de Congolais. De la même manière, la population Pélézi s'est avérée être différente des autres populations guinéennes alors même que sa proximité géographique avec ces autres populations est importante. La conclusion d'absence de corrélation, à l'échelle du pays, entre distance génétique et distance physique peut-être reliée aux résultats de Berthaud (1986) qui concluait en une absence de structure génétique au niveau de la Côte d'Ivoire. Nous avons montré qu'il existe en réalité une structure fine en populations mais avec des échanges *a priori* possibles. Il est important de noter que les populations étudiées ont été échantillonnées dans des forêts résiduelles et qu'elles sont isolées les unes des autres. On peut émettre l'hypothèse qu'une structure de population telle que celle que l'on a décrite en Ouganda (voir Annexe A.3.5) peut avoir existé, permettant des échanges entre populations par exemple grâce à des arbres relais, mais une telle dynamique ne peut plus exister aujourd'hui, du fait de la fragmentation des reliquats de forêts primaires. Une étude plus approfondie et plus systématique des populations résiduelles de caféiers spontanés en Côte d'Ivoire pourrait permettre de mieux comprendre la dynamique des échanges de gènes entre les différentes populations.

D'après les résultats présentés dans la seconde partie de ce chapitre, il semble que les 15 marqueurs que nous avons utilisés seraient suffisants pour étudier l'ensemble des collections de manière globale et rapide. Néanmoins la structure très fine à l'échelle des différentes populations pour des origines à base génétique plus large nécessitera un plus grand nombre de marqueurs, comme nous le montrerons dans le chapitre suivant. Un certain nombre d'origines restent à caractériser, notamment les génotypes du Cameroun qui, comme nous l'avons abordé dans la première partie de ce chapitre, sont vraisemblablement proches ou appartiennent au groupe C. Une étude plus poussée des génotypes provenant du bassin du Congo serait également souhaitable, cette région semblant renfermer une importante diversité. Enfin des prospections dans la dernière zone où existent des génotypes de *C. canephora* à l'état sauvage non étudiés, au nord de l'Angola, pourra finir de préciser la connaissance de la

diversité génétique de cette espèce. Néanmoins des données sur d'autres espèces, comme le palmier à huile *Elaeis guineensis* (Cochard, communication personnelle), ainsi que les propos de Cramer (1957) semblent indiquer que cette origine sera proche ou partie du groupe SG1.

Les origines cultivées que nous avons étudiées, les groupes SG1 et SG2, ne correspondent qu'à une partie restreinte de la diversité disponible de l'espèce. Cet état de fait, déjà avancé par Montagnon *et al.* (1998b) permet d'envisager d'utiliser l'ensemble des ressources génétiques de notre espèce pour l'amélioration. Pour cela, des compléments de caractérisation des différents groupes mis en évidence pour quelques caractères impliqués dans l'élaboration de la qualité du café à la tasse (taux de caféine, saccharose,...) et dans les caractères agronomiques (production, granulométrie (taille des grains), résistance à la sécheresse,...) pourront être menés. Cette caractérisation pourra se faire dans le cadre d'études d'association qui permettront l'identification des zones du génome impliquées dans les variations de ces caractères. Des études préliminaires portant sur l'étendue du déséquilibre de liaison, préalable nécessaire à la mise en place d'études d'association, seront menées au sein de différentes populations et feront l'objet du prochain chapitre.

Enfin les données de cartographie génétique et les zones QTL déjà mises en évidence pourront donner une base pour le choix de régions potentiellement intéressantes à explorer avant de pouvoir généraliser ces approches à l'ensemble du génome. L'un des principaux intérêts de la génétique d'association étant en effet la validation et la précision de zones QTL déjà connues.

Un résultat à court terme d'une telle étude de diversité serait aussi d'éviter les redondances de génotypes dans les collections et de permettre l'établissement de core-collections. Celles-ci permettent à la fois une meilleure gestion des ressources génétiques et la possibilité d'établir des collections de travail en réduisant le nombre de génotypes à conserver et à phénotyper, ainsi que les coûts importants inhérents à la mise en place de telles collections sur des espèces pérennes. La connaissance de la structure de ces core-collections, et sa maîtrise, permettraient de les utiliser également à des fins de génétique d'associations. Au vu de l'importante structure existant entre les Guinéens et les Congolais, il semblerait cohérent de réaliser trois types de core-collections, une pour chacun de ces deux grands groupes de diversité (Congolais et Guinéens), et une au niveau global de l'espèce.

Chapitre 4 : Les patrons de déséquilibre de liaison au sein de quelques groupes de *Coffea canephora* étudiés à l'aide de microsatellites.

Introduction :

La connaissance de l'importance du déséquilibre de liaison, de son étendue, de sa répartition au sein de l'espèce étudiée est un préalable indispensable à des approches de génétique d'associations (Zhu *et al.*, 2008). Les études d'association sont particulièrement pertinentes pour les espèces pérennes car elles peuvent être réalisées sur des populations préexistantes, soit naturelles soit d'amélioration, et ne nécessitent donc pas la création de populations spécifiques par croisements contrôlés comme dans le cas des approches classiques de cartographie génétique. L'analyse de ces populations naturelles ou d'amélioration permet un gain de temps précieux pour détecter les associations entre des caractères phénotypiques et des marqueurs moléculaires par rapport aux approches de cartographie génétique.

De telles approches ont montré, ou montrent, des preuves d'efficacité chez un nombre croissant d'espèces végétales, citons parmi celles-ci le maïs (Yu & Buckler, 2006), la vigne (This, communication personnelle), le pin (Ingvarsson, 2005; Gonzalez-Martinez *et al.*, 2007; Gonzalez-Martinez *et al.*, 2008). L'étude préalable du DL au sein de l'espèce oriente le choix d'une des 2 grandes approches d'étude d'association : « Scan Génome Entier » ou « Approche Gène Candidat ».

Notre travail a donc été d'évaluer le DL au sein de divers groupes de *C. canephora* que nous avions précédemment identifiés :

- 2 populations naturelles, Pélézi et Nana (groupe C)
- un groupe composite de génotypes d'origine ivoirienne et guinéenne (groupe G)
- un groupe composite de génotypes cultivés originaire du bassin du Congo (groupe SG2)

- un groupe composite de populations originaires de la façade atlantique du Gabon, du Congo Brazzaville ou de la RDC (groupe SG1).

Les résultats attendus de cette partie de notre travail sont i) une meilleure connaissance de la structure génétique fine de notre espèce ii) la connaissance de la dynamique du DL au niveau génomique pour un certain nombre de groupes de diversité ou de populations iii) l'identification de populations adaptées à l'une ou l'autre approche de la génétique d'association.

Matériel et méthodes

Matériel végétal

Le matériel végétal étudié dans cette partie représente un total de 356 génotypes répartis en 5 populations ou groupes de diversité en se basant sur les analyses de diversité du chapitre précédent :

- SG2, génotypes cultivés en Côte d'Ivoire et en Ouganda ou en collection au Brésil et vraisemblablement originaires du bassin du Congo
- SG1, génotypes cultivés au Togo et au Bénin ou issus de prospection en collection à Luki en RDC (Bieysse, communication personnelle) et originaires de la façade atlantique allant du Gabon au Congo. La population Luki, proche des Niaouli de ce groupe (chapitre précédent), a en effet été intégrée à l'étude des SG1
- Nana, groupe C, génotypes issus de prospections en forêt en République de Centre Afrique. Un certain nombre de génotypes appartenant à cette population qui n'avaient pas été caractérisés dans l'étude générale de la diversité effectuée au chapitre précédent ont été inclus afin de porter l'effectif de cette origine à un total de 92 génotypes.
- Pélézi, population naturelle et isolée de Guinéens de Côte d'Ivoire
- Guinéens, mélange de différentes populations sauvages ou cultivées de Côte d'Ivoire et de Guinée.

Nous n'avons pas conservé pour ces analyses la population Libengé dont l'effectif nous paraissait faible pour établir des conclusions générale sur le DL dans cette population.

L'échantillonnage complet du matériel végétal est donné en Annexe A.4.1.

Choix des marqueurs microsatellites et génotypage

Un total de 108 marqueurs microsatellites d'origines diverses a été génotypé sur l'ensemble des individus. Ces marqueurs sont issus de différentes origines que nous allons décrire ci-après. L'ensemble des marqueurs utilisés dans cette étude a été cartographié sur une carte génétique intraspécifique développée au CIRAD et qui comprend à ce jour 249 marqueurs (microsatellites et fragments de gènes candidats) pour 248 individus. Pour rappel cette carte est composée de 11 groupes de liaison et couvre un total de 1407,6 cM.

L'ensemble des 11 groupes de liaison a été étudié, avec un nombre plus important de marqueurs sur les groupes de liaison A, B, D, F, G et H. Une localisation sur la carte génétique des marqueurs utilisés est donnée en Figure 4.1.

Cette approche, avec en moyenne un marqueur tous les 13 cM mais des zones fortement densifiées, nous semblait être la stratégie la mieux adaptée pour construire une première vision pan-génomique du DL chez notre espèce tout en permettant la comparaison du comportement du DL sur des groupes de liaison différents. Ceci devrait permettre une meilleure appréhension des possibilités d'études d'association au sein de nos populations et guider la stratégie de mise en place de telles études dans deux directions :

- Premièrement le choix des populations d'association, et le type d'étude potentielle sur ces populations
- Deuxièmement le nombre de marqueurs à utiliser dans de telles études avec notamment la densité envisagée pour permettre de capter le plus grand nombre possible d'associations potentielles.

Marqueurs issus de banques enrichies en motifs microsatellites

Les marqueurs notés en 3 chiffres sont issus de 2 banques enrichies en motifs microsatellites, l'une développée au CIRAD (Poncet *et al.*, 2007) à partir du clone 126 de *C. canephora*, la seconde développée à l'IRD à partir d'une variété de *C. arabica*, la variété Cattura (Combes *et al.*, 2000). Ces marqueurs ont été testés pour leur transférabilité sur les différentes espèces du genre *Coffea* (Combes *et al.*, 2000; Poncet *et al.*, 2007; Cubry *et al.*, 2008).

Les marqueurs notés ssrXxxx sont eux aussi dérivés d'une banque enrichie en motifs microsatellites et nous ont été fournis par le centre de recherche Nestlé de Tours.

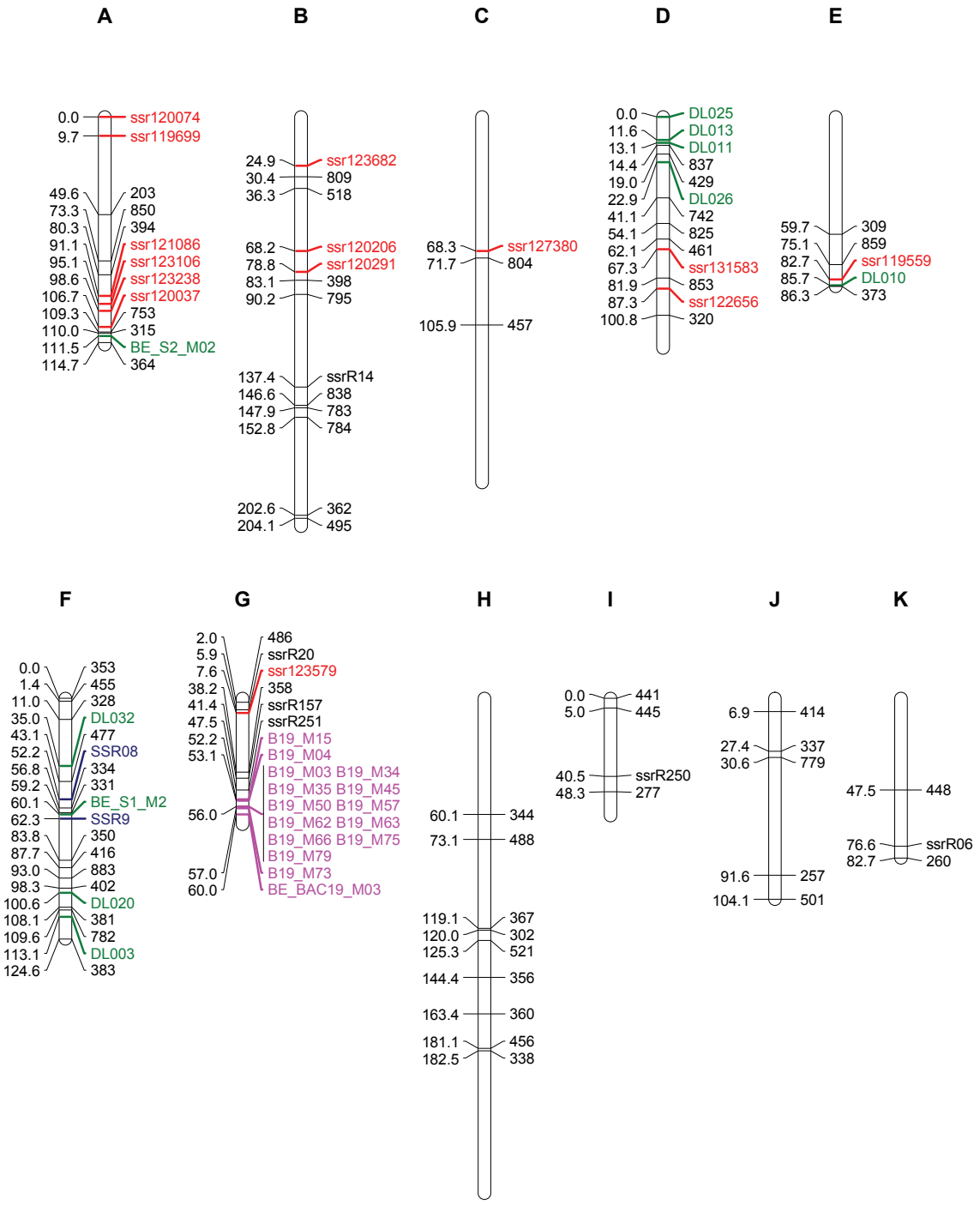


Figure 4.1 : Localisation sur la carte génétique des 108 marqueurs microsatellites utilisés pour l'étude de déséquilibre de liaison. En noir les marqueurs issus de banques génomiques, en rouge les marqueurs dérivés d'EST, en bleu les marqueurs identifiés sur des séquences du gène Susy et de son promoteur, en vert les marqueurs développés sur des séquences de Bac-ends et en violet les marqueurs développés sur la séquence du Bac 111O18

Marqueurs issus de séquençage d'une banque EST

Les marqueurs ssrxxxxxx sont dérivés de séquences d'ADNc issus d'une banque d'EST développée par Nestlé. Ces marqueurs nous ont également été fournis par le centre de recherche Nestlé de Tours. Ils permettront de mener des études de cartographie comparée entre les cartes développées par les différentes équipes.

Marqueurs issus de séquences de gènes candidats et de BAC-ends de clone BACs ayant hybridé avec des séquences de gènes candidats

Les marqueurs SSR08 et SSR09 sont dérivés de motifs microsatellites identifiés dans des séquences promotrices d'un gène candidat de Saccharose Synthétase (Susy1).

Les marqueurs DLxxx et BE-xxx ont été développés dans des séquences de BAC-ends de BAC s'hybridant avec des sondes de gènes candidats (Susy1 et 2, Cell Wall Invertase) et de sondes génomiques (gA71, gA10).

Les microsatellites B19-Mxx ont été développés à partir de la séquence consensus d'un clone BAC (111O18) orthologue du BAC19 de la tomate (*Solanum lycopersicum*), séquencé en collaboration avec Nestlé dans le cadre d'une étude de synténie Caféier/Tomate (Crouzillat, communication personnelle). Une analyse plus approfondie du déséquilibre de liaison dans cette région du génome est l'objet du prochain chapitre. Pour ces marqueurs, nous avons utilisé des distances génétiques approximatives, seulement certains marqueurs étant cartographiés. Connaissant la séquence de ce clone BAC, nous avons établi ces distances selon l'organisation des marqueurs sur la séquence et par rapport à des marqueurs cartographiés.

Génotypage et acquisition des données

Le génotypage a été réalisé en suivant le protocole décrit dans le premier article issu de cette thèse (Cubry *et al.*, 2008). Des témoins de taille on été répétés sur les différents gels afin de permettre une homogénéité des données de génotypage. Les données ont été importées à partir de SAGA GT® et formatées pour le logiciel PHASE (Stephens *et al.*, 2001) à l'aide du logiciel CREATE (Coombs *et al.*, 2007). Les sorties de PHASE ont ensuite été formatées pour l'analyse de diversité (DARwin 5.0) et l'analyse du DL (PowerMarker).

Analyses statistiques des données et calcul du déséquilibre de liaison

Reconstruction des haplotypes

L'espèce *C. canephora* est une espèce fortement hétérozygote comme l'ont montré les analyses précédentes, il est donc impossible de distinguer la phase des allèles des doubles hétérozygotes Aa/Bb, c'est-à-dire de savoir si A est associé avec B ou avec b au niveau haplotypique. Or les mesures les plus courantes et les plus puissantes du déséquilibre de liaison (*D*, *D'*, r^2) reposent sur l'estimation d'un déséquilibre de liaison haplotypique ou gamétophytique, c'est-à-dire faisant appel à la phase des allèles au niveau des gamètes. Afin de permettre une estimation de ces mesures, il nous faut donc avoir accès aux haplotypes. L'accès aux haplotypes pour des espèces diploïdes hétérozygotes peut être réalisé soit par des techniques de laboratoires lourdes telles que le clonage, soit par des modèles de génétique des populations. Nous avons utilisé le logiciel PHASE (Stephens *et al.*, 2001; Stephens & Donnelly, 2003; Stephens & Scheet, 2005) afin de reconstruire ces haplotypes. Ce logiciel estime les haplotypes les plus probables pour chaque génotype en se basant sur un algorithme EM (Expectation-Maximisation) incorporant une hypothèse de coalescence dans un modèle à maximum de vraisemblance. Les haplotypes sont donc reconstruits suivant un certain nombre d'hypothèses fortes à l'aide de paramètres tels que les fréquences alléliques, les possibilités de recombinaison entre les marqueurs et les généalogies simulées des allèles.

Nous avons utilisé l'ensemble des 356 génotypes car l'algorithme fonctionne de manière plus satisfaisante sur des données structurées (Stephens & Donnelly, 2003). Le jeu de données a néanmoins été partitionné par Groupe de Liaison de la carte génétique (soit 11 matrices au total), les marqueurs situés sur des groupes différents ne pouvant être en phase. Cette approche a montré sa pertinence et le gain de puissance dans l'évaluation du déséquilibre de liaison est avéré dans le cas de la vigne (Barnaud *et al.*, 2006).

Cinq répétitions de l'algorithme, basées sur 1000 itérations, 100 thining-intervals et 1000 burn-in ont été réalisées. La répétition montrant le plus fort maximum de vraisemblance a été conservée pour la suite des analyses.

Analyse de la diversité

Nous avons réalisé pour l'ensemble des individus et pour les différents groupes des Analyse Factorielle sur Tableau de Dissimilarités (AFTD) à l'aide du logiciel DARwin 5.0 (Perrier *et al.*, 2003; Perrier & Jacquemoud-Collet, 2006) afin de valider la structure

précédemment décrite et d'incorporer les nouveaux génotypes (Nana et Luki) à cette diversité. Nous avons également étudié plus précisément la structure du groupe des Guinéens. Afin de compléter cette analyse, nous avons généré des arbres de diversité à l'aide de l'algorithme de Neighbour-Joining pour l'ensemble des groupes étudiés.

Analyse du déséquilibre de liaison

Les 5 groupes identifiés ont été analysés conjointement et séparément pour le déséquilibre de liaison. Plusieurs tests ou statistiques ont été utilisés pour cette analyse.

Tests exact d'association entre paires de marqueurs

Afin de tester si les fréquences haplotypiques entre 2 locus sont le produit des fréquences alléliques correspondantes aux 2 locus, des tests exacts de Fisher ont été réalisés pour l'ensemble des combinaisons possibles. Le comptage des allèles est organisé en tableau de contingence et des permutations suivant un algorithme utilisant une chaine de Markov-Monte-Carlo sont utilisées pour calculer les p-values non biaisées associées au test (Weir, 1996; Liu & Muse, 2005). Nous avons corrigé le seuil de signification des p-values associées au test exact à l'aide de la correction de Bonferroni afin de s'affranchir au mieux de l'effet du nombre très important de tests réalisés, et ce bien que cette correction soit très conservatrice.

L'ensemble des associations possibles a été testé. Le nombre d'associations significatives a été évalué en intra et intergroupes de liaison, et un ratio du nombre d'associations significatives entre les associations intragroupe et intergroupe a été calculé.

Mesures du déséquilibre de liaison

Nous avons calculé les valeurs de D' et r^2 pour l'ensemble des combinaisons possibles de marqueurs 2 à 2 à l'aide du logiciel PowerMarker. Ces mesures ont initialement été développées pour des locus bialléliques. Néanmoins une estimation de ces mesures pour des marqueurs multialléliques est réalisée en faisant une moyenne pondérée de l'ensemble des déséquilibres entre paires d'allèles (encadré 4.1).

Encadré 4.1 : Extensions multialléliques des mesures de DL

Les mesures classiques du Déséquilibre de liaison (*D'* et r²) ont été fondamentalement décrites pour des locus bi-alléliques. Les microsatellites étant généralement multialléliques, des adaptations de ces mesures doivent être utilisées pour comparer les locus 2 à 2 sur l'ensemble du génome.

$$D'_m = \sum_{u=1}^{a_v} \sum_{v=1}^{b_v} p_u p_v \left| D'_{AuBv} \right| \quad \text{(Liu \& Muse, 2005)}$$

$$r^2_m = \sum_{u=1}^{a_v} \sum_{v=1}^{b_v} p_u p_v \left| r^2_{AuBv} \right| \quad \text{(Liu \& Muse, 2005)}$$

Ces mesures sont des moyennes pondérées des déséquilibre calculés pour chaque paire d'allèles possible entre 2 locus. Elles sont implémentées dans les logiciels PowerMarker et TASSEL.

Résumé des résultats et représentation graphique

Au vu du nombre très important d'associations testées, des représentations graphiques sont indispensables pour résumer au mieux les résultats obtenus. Nous avons représenté les valeurs de r² en fonction de la distance pour chaque groupe ou population étudié. De la même manière, des matrices 2 à 2 des valeurs de r² et des p-values associées aux tests exacts ont été réalisées. *D'* n'a été utilisé que pour l'ensemble des génotypes.

Résultats

Reconstruction des haplotypes

Afin de permettre l'utilisation de PHASE en conservant l'hypothèse de mutation « pas à pas » (Stepwise Mutation Model) pour les marqueurs microsatellites, une transformation de nos données, exprimées en tailles d'allèles, en nombre de répétitions a été réalisée à l'aide du logiciel CREATE (Coombs *et al.*, 2007).

Les différents haplotypes ont été reconstruits pour chaque groupe de liaison. Pour chaque jeu de données, la répétition du modèle ayant obtenu la valeur maximale de vraisemblance a été retenue. Les matrices ont ensuite été fusionnées et formatées pour les

incorporer dans PowerMarker afin d'analyser le DL, en déclarant comme type de données des données génotypiques à phase connue.

Validation de la structure génétique

Ensemble des génotypes

L'AFTD sur les 4 premiers axes et l'arbre correspondant sont donnés en Figure 4.2. Les nouveaux génotypes incorporés se placent dans les groupes de diversité préalablement identifiés. Les axes 1 et 2 représentent respectivement 29,9 et 17,9% de la variabilité totale et séparent les groupes Guinéens et Congolais (axe 1) et le sous-groupe Nana des autres Congolais (axe 2). De même l'axe 3 sépare nettement les SG1 alors que l'axe 4 sépare les Pélézi des autres Guinéens (non montré) bien que leurs explications respectives soient restreintes (4,59% et 2,96%). Cette analyse permet par ailleurs d'identifier 3 individus intermédiaires qui correspondent vraisemblablement à des hybrides entre les groupes de diversité identifiés. Ces individus (G5011, C2007 et G1011) se placent en effet entre les groupes SG2 et G. Nous validons la structure à priori en 5 groupes précédemment énoncée. Ce résultat est confirmé par l'arbre NJ associé, représentant l'ensemble des génotypes sauf les trois identifiés comme hybrides.

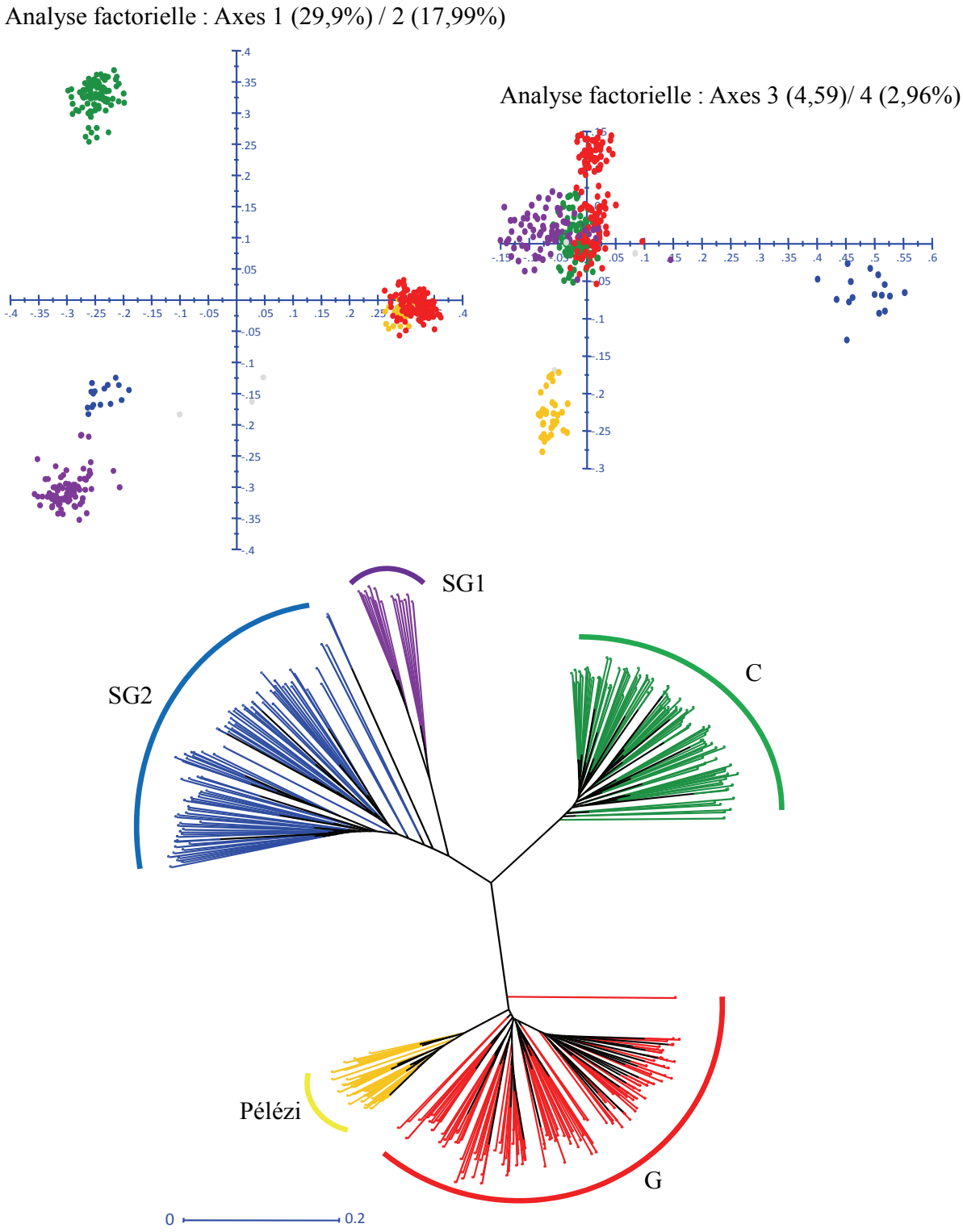


Figure 4.2 : 4 premiers axes de l'AFTD (Analyse Factorielle sur Tableau de Dissimilarités) pour l'ensemble des 356 individus et arbre de Neighbour-Joining sur 353 individus, les 3 génotypes identifiés comme hybrides sur l'AFTD (en gris) ayant été enlevés. L'échelle de distance indiquée est basée sur l'index de dissimilarité Simple-Matching calculé à l'aide de DARwin (Perrier *et al.,* 2003)

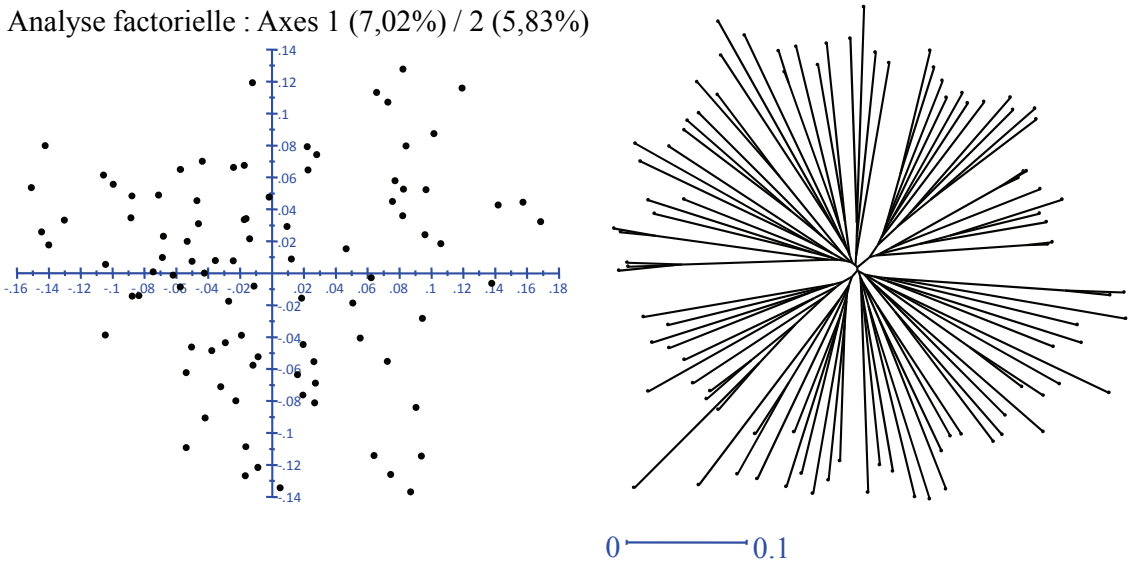
Groupe de diversité C et Pélézi (groupes-populations)

Les 2 premiers axes des AFTD et les arbres Neighbour-Joining associés pour ces 2 groupes sont donnés en Figure 4.3.

Pour les Nana et les génotypes associés (groupe C), l'AFTD ne montre pas de structure avec un nuage d'individus bien réparti sur l'ensemble des 2 axes principaux. Les axes 3 et 4 ne montrent pas non plus de structure (non montré). De plus l'explication des 2 premiers axes est relativement faible, indiquant une faible part de la variabilité due à une quelconque structure. L'arbre de diversité associé se présente comme une étoile, ce vers quoi on s'approche en cas d'absence de structure ou après ré-échantillonnage (par exemple dans le cas de core-collection) réalisé pour briser la structure (Perrier, communication personnelle). On s'approche donc *a priori* pour cette population d'un cas intéressant avec un DL potentiellement dû uniquement à des effets démographiques ou de mutations, sans interférence de la structure. Nous conserverons donc l'ensemble des génotypes de ce groupe, bien que les données de prospection indiquent des niveaux d'apparentements divers, avec une cinquantaine de génotypes transférés par bouture à partir de la population d'origine, alors que les autres génotypes collectés ont été prélevés sous forme de graines sur 6 pieds mères différents, formant donc 6 familles de demi-frères (Berthaud *et al.*, 1984).

En ce qui concerne la population Pélézi, on obtient un nuage relativement homogène sur les 2 premiers axes. Néanmoins le pourcentage de variabilité expliqué par ses axes est élevé indiquant une part non négligeable de cette variabilité due à la répartition des génotypes sur ces 2 axes. L'analyse de l'arbre correspondant aboutit à la même conclusion d'une population relativement homogène génétiquement mais avec une certaine hiérarchisation, sans doute due à plusieurs niveaux d'apparentements. En effet cette population correspond à la prospection de quelques arbres mères et d'un certain nombre de juvéniles (Leroy, communication personnelle). Les branches internes de l'arbre restent courtes, avec une faible diversité génétique présente, ce qui peut permettre de dire que seul des effets d'apparentement risquent d'interférer sur le calcul de DL. Nous ne pouvons par ailleurs pas identifier de sous-structure évidente au sein de cette population, ce qui nous amènera à la considérer dans son ensemble pour les analyses de DL.

Groupe C



Groupe Pélézi

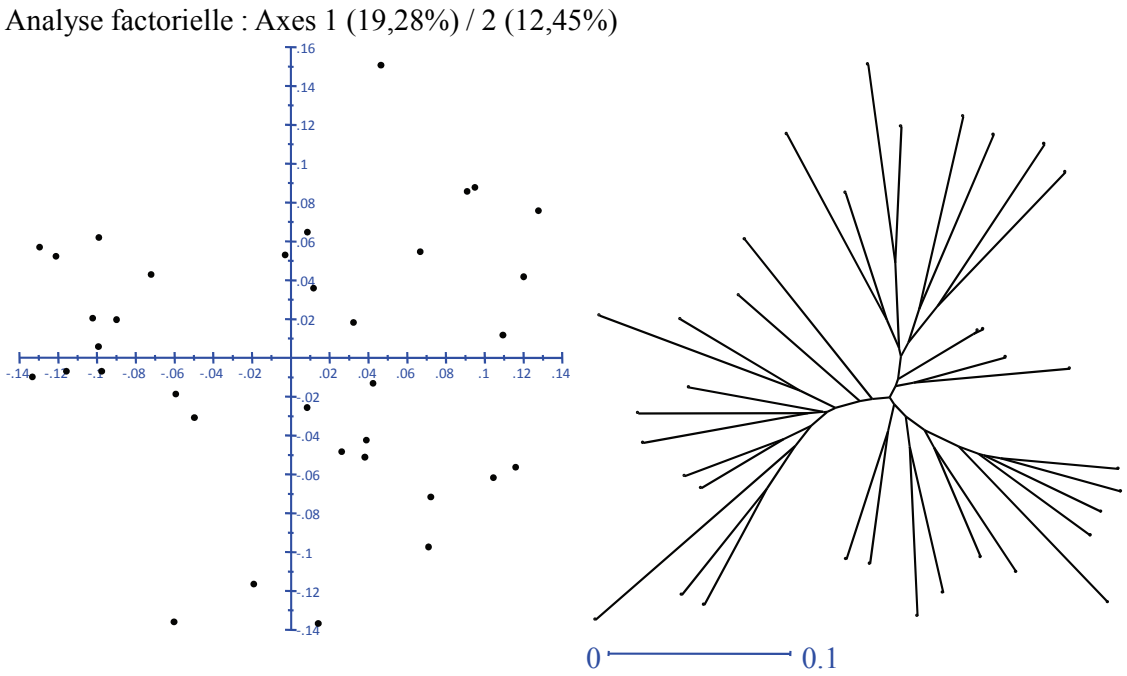


Figure 4.3 : 2 premiers axes de l'AFTD et arbres NJ associés pour les 2 groupes-populations C (en haut) et Pélézi (en bas). Les pourcentages d'explication de la variabilité totale sont donnés pour chaque axe. Les échelles de distances correspondent à l'index de Simple-Matching

Origines cultivées (SG2)

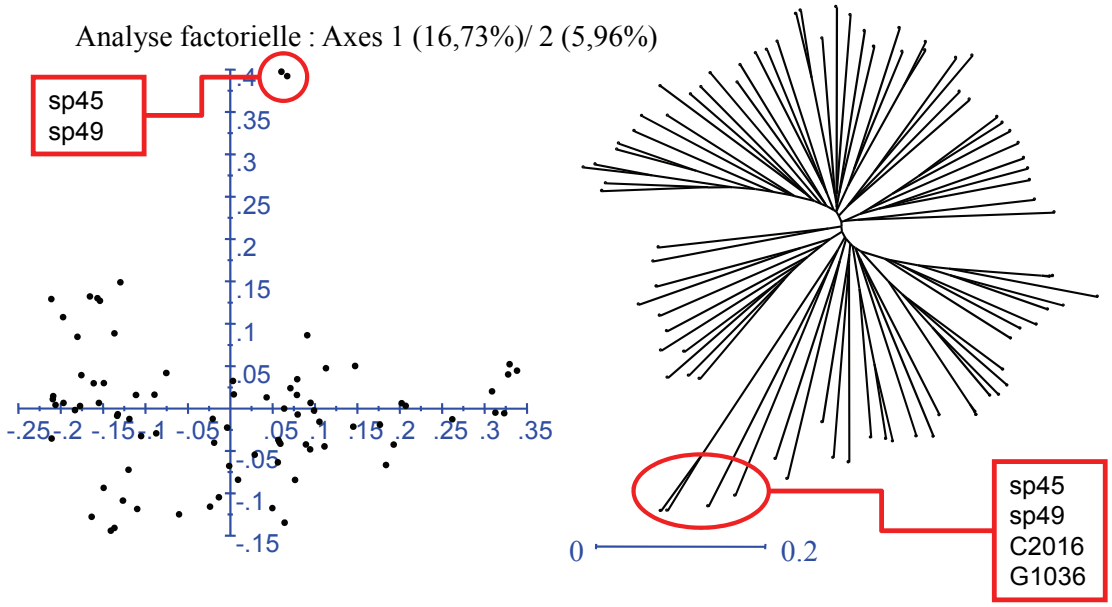
Les SG2 (Figure 4.4) ne semblent pas montrer de structure génétique évidente mais 4 génotypes sont plus distants (sp45 et sp49, C2016 et G1036) sur l'arbre, dont 2 sont également nettement séparés du reste du groupe sur l'AFTD. Nous écarterons donc ces individus de l'analyse de DL afin d'éviter un effet de structure dû à ceux-ci. Une fois les 4 génotypes écartés, on constate que l'AFTD ne permet plus de discerner des génotypes plus distants et que l'arbre prend une forme étoilée, synonyme d'un échantillonnage peu ou pas structuré.

Groupes SG1 et G

Les AFTD et arbres pour ces deux groupes sont donnés en Figure 4.5. Le groupe SG1 est clairement structuré en 2 clusters tant sur l'AFTD que sur l'arbre associé. Cette structure n'est pas étonnante étant donnée l'origine diverse des génotypes étudiés, mais au vu du faible nombre de génotypes de ce groupe nous le conserverons en l'état pour l'analyse de DL, les résultats obtenus devront néanmoins être nuancés compte tenu de la structure sous jacente. Les deux clusters mis en évidence correspondent pour l'un à la population Niaouli qui comprend des génotypes cultivés, pour l'autre à la population Luki composée de génotypes spontanés.

Pour les Guinéens, l'AFTD semble faire apparaître une structure marginale en 3 clusters. Sur l'arbre ces 3 clusters correspondent à la population Mouniandougou plus quelques individus Piné et Guinéens Cultivés (sub1), à la majorité de la population Ira2 (sub2) et aux populations Fourougbankoro et Ira1 plus quelques individus Guinéens Cultivés et Piné (sub3). Cette étude permet de mieux préciser la structure génétique de ce groupe que l'étude précédente. Il semble qu'une structure en populations existe mais que celles-ci sont proches génétiquement et peu différenciées, confirmant notre étude précédente. Nous confirmons donc les résultats du précédent chapitre tout en les précisant et nous montrons que cette structure génétique très fine ne peut être étudiée qu'avec un nombre très important de marqueurs. Pour la suite de notre travail nous considérerons donc le groupe Guinéen comme un groupe composite de différentes populations et nous essaierons également d'analyser séparément les 3 sous-groupes que nous avons identifiés.

Groupe SG2 (origines cultivées)



Groupe SG2 moins sp45, sp49, C2016 et G1036

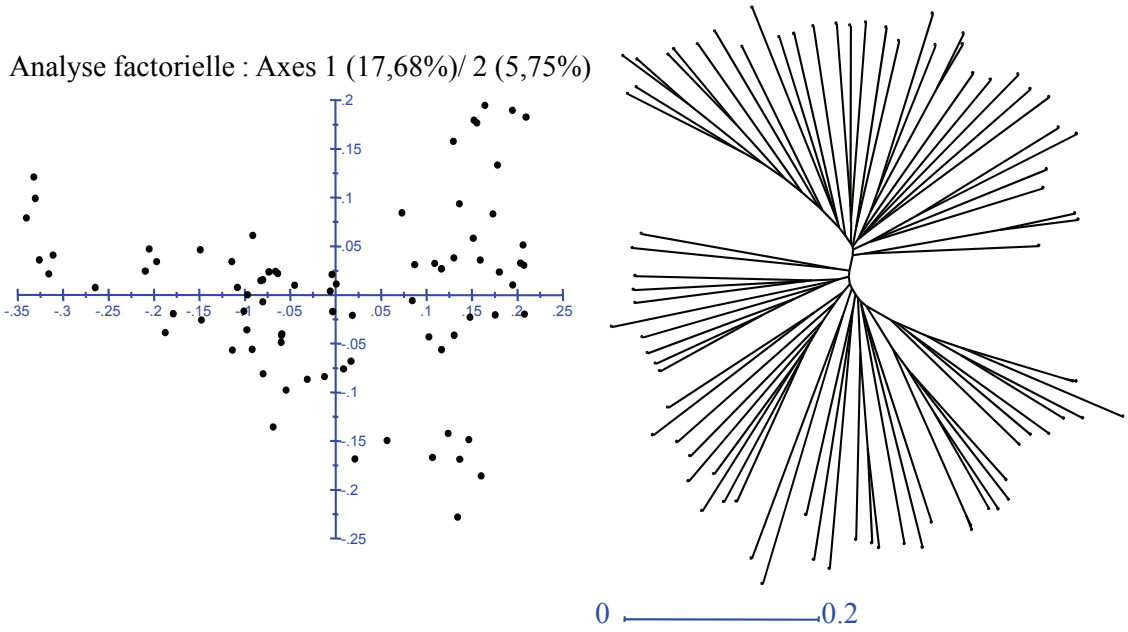


Figure 4.4 : 2 premiers axes de l'AFTD et arbres NJ associés pour le groupe SG2 avec tous les individus (en haut) et avec 4 génotypes légèrement distant enlevés (en bas). Les pourcentages d'explication de la variabilité totale sont donnés pour chaque axe. Les échelles de distance correspondent à l'index de Simple-Matching

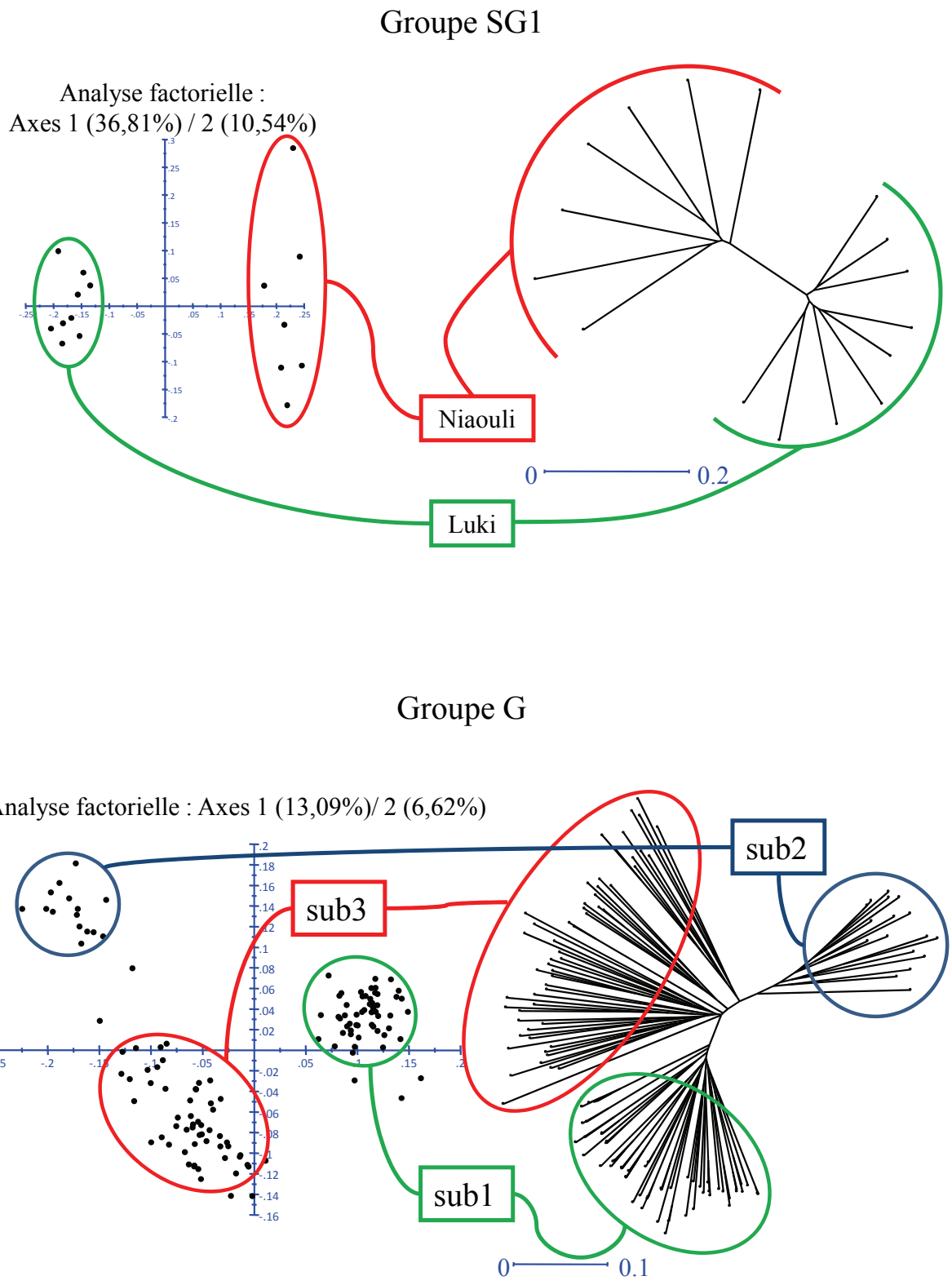


Figure 4.5 : 2 premiers axes de l'AFTD et arbres NJ associés pour les groupes SG1 (en haut) et G (en bas). Les pourcentages d'explication de la variabilité totale sont donnés pour chaque axe. Des sous-groupes cohérents ont été identifiés. Les échelles de distance correspondent à l'index de Simple-Matching

Analyse du DL au niveau pangénomique

Des analyses de DL ont été réalisées pour l'ensemble des génotypes, les 5 groupes identifiés et 3 sous-groupes de Guinéens.

Calcul des associations significatives intra et intergroupes de liaison par tests exacts de Fisher et correction de Bonferroni

Le tableau suivant présente un résumé des tests exacts effectués sur les différents clusters considérés (Tableau 4.1).

Tableau 4.1 : résumé des analyses d'association entre marqueurs par test exact de Fisher pour les différents groupes considérés.

Population	Nombre individus	Nombre de marqueurs polymorphes	Nombre de tests 2 à 2	Seuil (5%) après correction de Bonferroni	Nombre d'associations significatives (test exact)		Nombre d'associations significatives intragroupes et pourcentage par rapport au total des associations significatives		Nombre d'associations significatives intergroupes et pourcentage par rapport au total des associations significatives		Ratio intragroupe/ intergroupe
Tous les génotypes	356	108	5778	8.65351E-06	5684	98%	680	***12%***	5004	***88%***	0.14
C	92	98	4753	1.05197E-05	72	2%	62	***86%***	10	***14%***	6.20
SG1	82	93	4278	1.16877E-05	177	4%	29	***16%***	148	***84%***	0.20
SG2	84	107	5671	8.81679E-06	389	7%	144	***37%***	245	***63%***	0.59
SG2 moins sp45, sp49, C2016 et G1036	80	105	5460	9.15751E-06	237	4%	117	***49%***	120	***51%***	0.98
Pélézi	35	74	2701	1.85117E-05	116	4%	85	***73%***	31	***27%***	2.74
G	128	97	4656	1.07388E-05	483	10%	99	***20%***	384	***80%***	0.26
G sub1	51	85	3570	1.40056E-05	73	2%	49	***67%***	24	***33%***	2.04
G sub2	16	73	2628	1.90259E-05	38	1%	32	***84%***	6	***16%***	5.33
G sub3	53	96	4560	1.09649E-05	36	1%	20	***56%***	16	***44%***	1.25

Ces résultats montrent l'importance des associations significatives générées par la structure génétique. En effet pour l'ensemble des génotypes, on peut remarquer que 98% des couples de marqueurs sont en déséquilibre, qu'ils soient ou non liés. D'autre part les clusters qui apparaissaient les moins structurés (C, Pélézi et les 3 sous groupes de G) sont les seuls à montrer un ratio intragroupe/intergroupe supérieur à 1. On note que lorsqu'on essaie de

considérer une structure fine, on arrive à corriger un certain nombre d'associations intergroupes, c'est par exemple le cas quand on enlève les 4 individus (sp45 et sp49, C2016 et G1036) du groupe SG2 qui apparaissaient comme étant légèrement différents du reste de SG2 dans les analyses factorielles, de même que pour les G avec une correction très importante des DL génomiques pour l'ensemble des sous-groupes vis-à-vis du groupe G. La prise en compte de la structure génétique dans les études d'association qui pourront être menées sur *C. canephora* sera donc indispensable pour éviter les fausses associations. Il semble de plus que le ratio du nombre d'associations significatives intragroupes de liaison sur le nombre d'associations significatives intergroupes de liaison pourrait être un premier estimateur grossier de la structure de nos populations.

Les valeurs résiduelles de DL génomique au niveau des groupes les moins structurés peuvent être expliquées par différents niveaux d'apparentement au sein des populations naturelles de notre espèce. En effet les populations naturelles de caféiers sont généralement de petites populations isolées avec un faible nombre d'arbres mères et quelques juvéniles, impliquant des relations de parenté importantes malgré l'allogamie stricte de notre espèce.

Décroissance du déséquilibre de liaison avec la distance génétique pour l'ensemble des génotypes

La décroissance du DL avec la distance est un phénomène généralement observé chez toutes les espèces. En espérance, dans le cas d'une population à l'équilibre entre mutation et dérive génétique, on s'attend à ce que le DL (mesuré par r^2) dépende à la fois de l'effectif efficace de la population et du taux de recombinaison entre les locus considérés. Le DL sera d'autant plus long à dissiper entre 2 marqueurs que ceux si sont proches. On pense donc trouver, à un instant donné de l'évolution d'une population, des DL plus important entre des marqueurs proches qu'entre des marqueurs éloignés. Nous avons représenté graphiquement le DL mesuré par r^2 et *D'* en fonction de la distance. Ces 2 valeurs ayant des propriétés différentes, les informations fournies par ces 2 paramètres ne sont pas redondantes. En effet *D'* ne va tenir compte que des recombinaisons alors que r^2 va aussi mesurer la mutation. Ces propriétés différentes font que *D'* sera globalement plus élevé que r^2 et pourra potentiellement mesurer plus d'associations, en revanche r^2 est plus informatif sur la manière dont seront associés les marqueurs et les caractères et est directement relié à la puissance potentielle des études d'association. De plus *D'* semble plus sensible aux faibles effectifs. Pour ces raisons on favorise généralement *D'* dans le cas de l'étude de l'histoire évolutive des populations

faisant appel à des panels importants d'individus. En revanche, dans l'optique d'études d'association, la valeur la plus indicative et la plus usitée chez les plantes est r^2.

Nous ne citerons les valeurs de *D'* qu'à titre de comparaison générale sur l'ensemble des individus avant de discuter plus en détail des différences de r^2 entre les différentes populations. Nous ne présenterons donc ici les valeurs de *D'* que pour l'ensemble des individus, les graphiques représentant les valeurs de *D'* en fonction de la distance pour les différents groupes seront donnés en Annexe A.4.2.

La Figure 4.6 représente les valeurs de r^2 et *D'* correspondant à des associations significatives après correction de Bonferroni en fonction de la distance pour l'ensemble des génotypes. Nous avons choisi de ne représenter ici que les associations significatives puisqu'elles représentent 98% des associations totales.

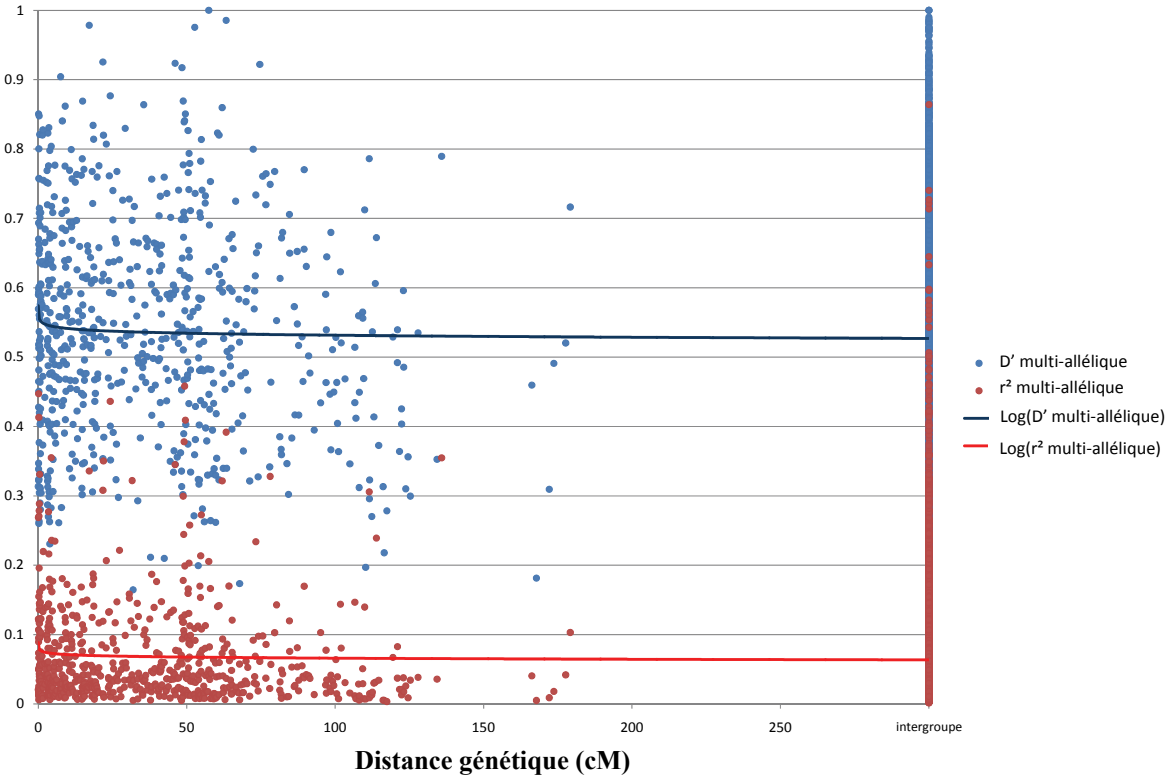


Figure 4.6 : *D'* et r^2 en fonction de la distance génétique en centiMorgan pour l'ensemble des individus. Seules les valeurs significatives après correction de Bonferroni sont représentées. Les courbes de régression logarithmique ont été calculées pour l'ensemble des données

De manière générale les valeurs de *D'* sont beaucoup plus élevées que les valeurs de r^2. De plus il semble que ces valeurs soient beaucoup plus stochastiques avec des proportions importantes de DL très élevés qui correspondent à des associations non significatives. Si l'on

s'intéresse aux différents groupes de diversité (figures données en Annexes A.4.2) *D'* semble très sensible aux effets de structure et est de plus sensible aux variations de fréquences alléliques entre les différent marqueurs. En conséquence, bien que les valeurs des 2 mesures apparaissent décroitre avec la distance il semble que *D'* soit moins sensible que r^2 à cette décroissance, avec des valeurs restant élevées à très longue distance et y compris entre marqueurs non situés sur le même groupe de liaison.

Décroissance du déséquilibre de liaison avec la distance pour les différents groupes de diversité

Les résultats de cette analyse montrent une décroissance générale du DL avec la distance. Néanmoins on trouve autant de cas différents que de groupes étudiés (Tableau 4.1).

Groupes SG1, SG2 et G

Pour le groupe SG1 (Figure 4.7), on trouve des valeurs de r^2 moyennes à élevées en comparaison des autres populations avec une forte proportion de DL intergroupe de liaison. La structure génétique importante en 2 populations permet d'expliquer de tels résultats. Au vu de ces résultats, il apparait difficile de choisir un seuil de r^2 qui serait utilisable en génétique d'association pour cette population et qui permettrait d'éviter les risques de fausses détection.

Pour le groupe SG2, on observe également des valeurs comparables de r^2 significatives entre des marqueurs non liés et entre marqueurs liés, même si ce phénomène est légèrement corrigé par l'écart de l'analyse de 4 génotypes plus distants. En revanche dans cette population les valeurs de r^2 sont extrêmement faibles même pour des marqueurs très proches. Ceci peut s'expliquer par l'origine même de cette population qui a dû subir un brassage génétique plus important que les autres origines car elle provient vraisemblablement d'un centre de diversité important de *C. canephora* et est issue de sélection. On peut émettre l'hypothèse pour ce groupe que l'obtention de valeurs comparables entre marqueurs liés et non liés n'est pas dû à un phénomène de structure mais plutôt à l'absence ou au très faible déséquilibre existant dans cette population. Par conséquent, il faudrait plus d'un marqueur microsatellite par cM dans cette population pour espérer couvrir l'ensemble du génome de manière convenable et pouvoir l'utiliser dans des études d'association.

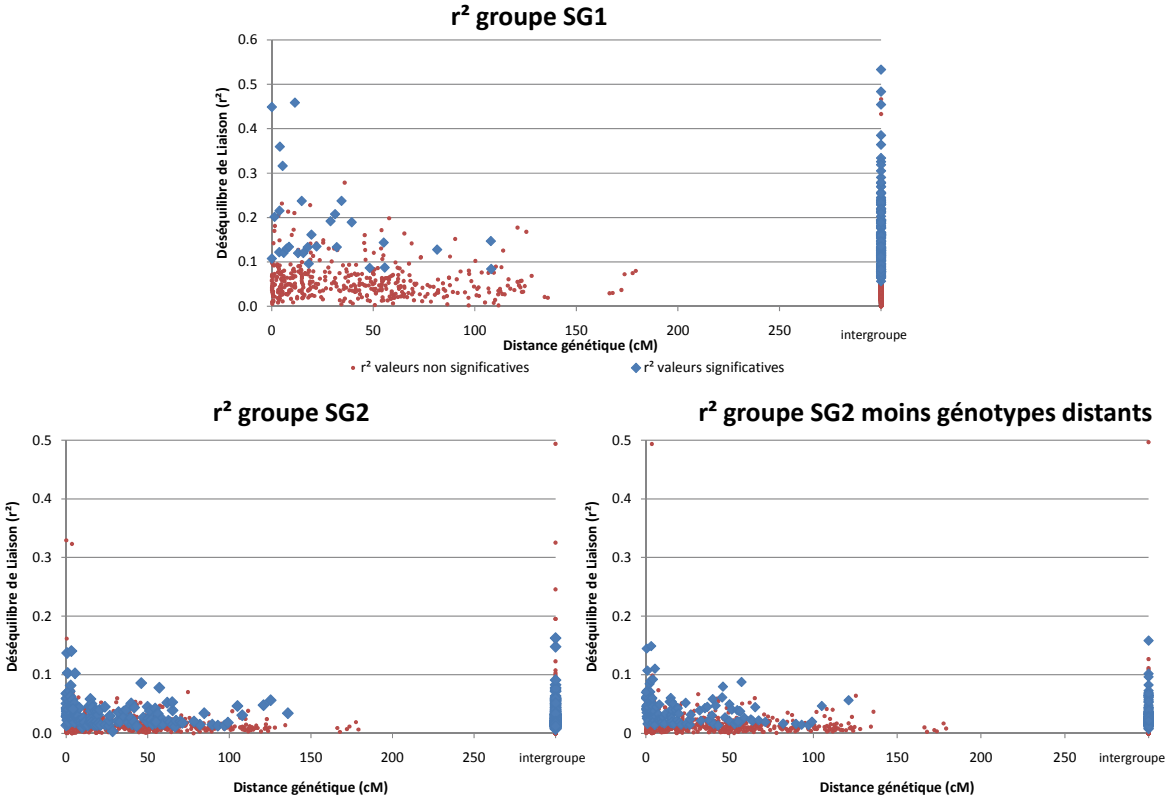


Figure 4.7 : r² en fonction de la distance génétique pour les groupes SG1 et SG2. Les régressions ont été calculées sur les valeurs significatives uniquement. En bleu les valeurs significatives, en rouge les valeurs non significatives après correction de Bonferroni du seuil de 5%

Le groupe G quant à lui présente des valeurs de r² plus élevées entre marqueurs liés qu'entre marqueurs non liés, avec une décroissance claire de ces valeurs avec la distance (Figure 4.8). Si pour l'ensemble du groupe quelques valeurs supérieures à 0,1 sont observables, cela n'est pas le cas pour les groupes Gsub1 et Gsub2. Dans ces 2 sous-groupes on observe effectivement des valeurs de r² significatives et élevées entre marqueurs proches, avec une décroissance prononcée. En choisissant un seuil empirique de r² de 0,2, qui semble raisonnable en vision des valeurs observées entre marqueurs non liés, on arrive à un DL couvrant environ 5 cM pour le groupe Gsub1 et 20 cM pour le groupe Gsub2. Ces valeurs permettent d'envisager couvrir l'ensemble du génome avec un nombre raisonnable de marqueurs, soit environ 280 et 70 marqueurs respectivement.

En ce qui concerne le groupe Gsub3 en revanche, on retrouve un cas proche de celui des SG2 avec *a priori* très peu de déséquilibre, même à faible distance, et même des valeurs qui apparaissent plus importantes entre des marqueurs non liés. Il faudrait donc

vraisemblablement plus de 1500 marqueurs pour couvrir l'ensemble du génome dans cette population. Néanmoins des effets de structure et d'apparentement pourraient expliquer ces résultats, le groupe Gsub3 apparaissant moins homogène que les 2 autres sous groupes sur les analyses de diversité.

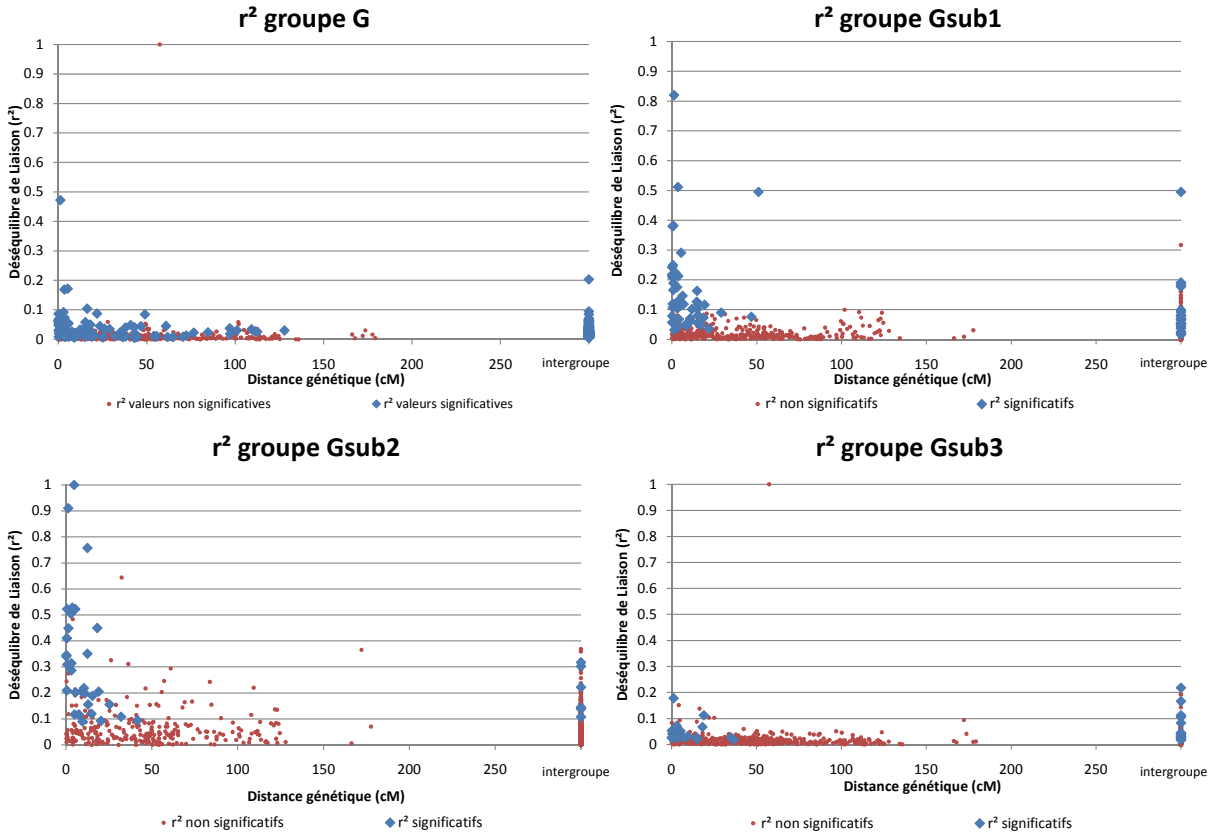


Figure 4.8 : r² en fonction de la distance génétique pour le groupe G et les sous groupes identifiés. Les régressions logarithmiques ont été calculées sur les valeurs significatives uniquement. Les valeurs en bleu sont les valeurs de r² significatives au seuil de 5% après correction de Bonferroni, les valeurs en rouge sont non significatives à ce seuil

Enfin les groupes C et Pélézi (Figure 4.9), nos 2 groupes-populations montrent des valeurs de r² élevées entre marqueurs proches et ayant une forte décroissance avec la distance, cette décroissance étant plus rapide chez les C que chez les Pélézi. Ce résultat permet d'envisager des potentialités intéressantes en génétique d'associations pour ces 2 populations. Si on choisit comme seuil, toujours en guidant notre choix par rapport aux valeurs observées entre marqueurs non liés, à 0,2 pour Pélézi et 0,1 pour C, on obtient une distance d'environ 23 cM pour Pélézi et 5 cM pour C. Ces distances permettent d'envisager des études d'association sur ces 2 populations avec respectivement 65 et 280 marqueurs environ pour Pélézi et C.

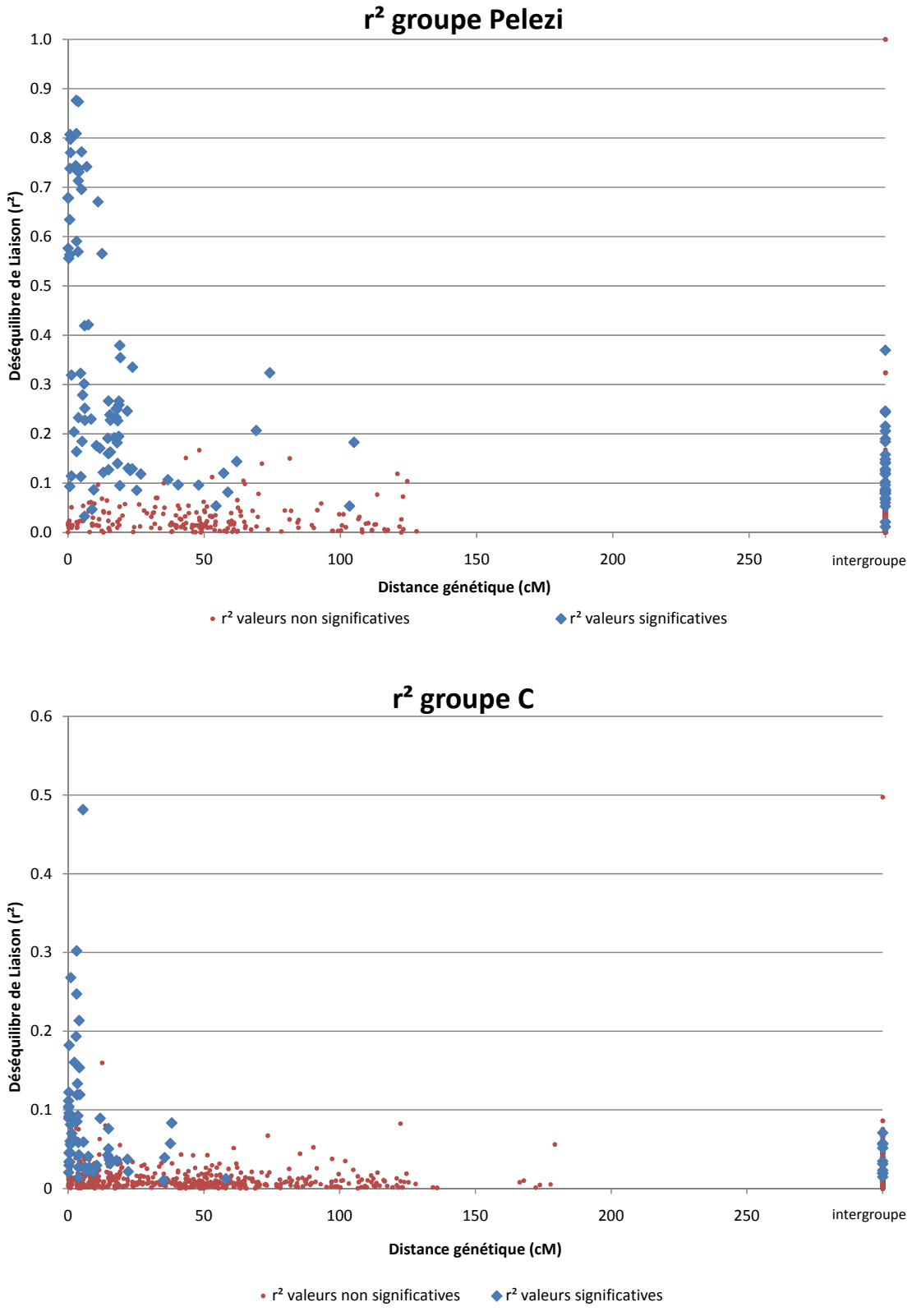


Figure 4.9 : r² en fonction de la distance génétique pour les groupes Pélézi (haut) et C (bas). Les valeurs en bleu sont les valeurs de r² significatives au seuil de 5% après correction de Bonferroni, les valeurs en rouge sont non significatives à ce seuil

La connaissance du DL global permet d'avoir une idée générale sur les variations de DL sur l'ensemble du génome. Cependant, le DL étant très variable suivant les régions, une analyse séparée de divers groupes de liaison pour les populations Pélézi, C, SG2 et G a été réalisée afin de comparer le DL suivant les régions du génome.

Comparaison des patrons de DL entre les groupes de liaisons

Les graphiques de r^2 en fonction de la distance pour les groupes de liaison A, B, D, F, G et H sont donnés en Figure 4.10 pour les groupes Pélézi, C, SG2 et G.

L'analyse de ces résultats permet de démontrer que le DL est variable en fonction des groupes de liaison considérés et des populations. Nous confirmons la quasi absence de valeurs utilisables de r^2 pour les groupes SG2 et G. Pour le groupe Pélézi, on observe des valeurs moyennes à élevées sur l'ensemble des groupes de liaison, avec néanmoins des différences importantes entre les groupes. Pour le groupe C, seuls les groupes de liaison A, F et G semblent montrer des valeurs de r^2 supérieures à 0,1. Cependant ces résultats sont à nuancer, notamment en ce qui concerne le groupe G qui possède une densité de marqueurs beaucoup plus importante que les autres groupes de liaison.

Si on observe les matrices de DL par population données en Figure 4.11, on observe que les p-values significatives se retrouvent préférentiellement proches de la diagonale donc entre marqueurs liés. Ceci est particulièrement vrai pour les groupes Pélézi et C, les groupes SG2 et G montrant une importante part de p-values significatives hors des groupes de liaison. La population C ne montre que quelques valeurs de r^2 supérieures à 0,1, reflétant la courte distance à laquelle le DL peut-être détecté. La population Pélézi, quant à elle, possède un nombre plus élevé de valeurs supérieures à 0,1 au sein des différents groupes de liaison avec un DL qui semble s'organiser en bloc. Ceci est concordant avec une hypothèse de régions ou le DL est important séparées par des régions de fortes recombinaisons ou le DL est beaucoup plus restreint (Flint-Garcia *et al.*, 2003; Rafalski & Morgante, 2004).

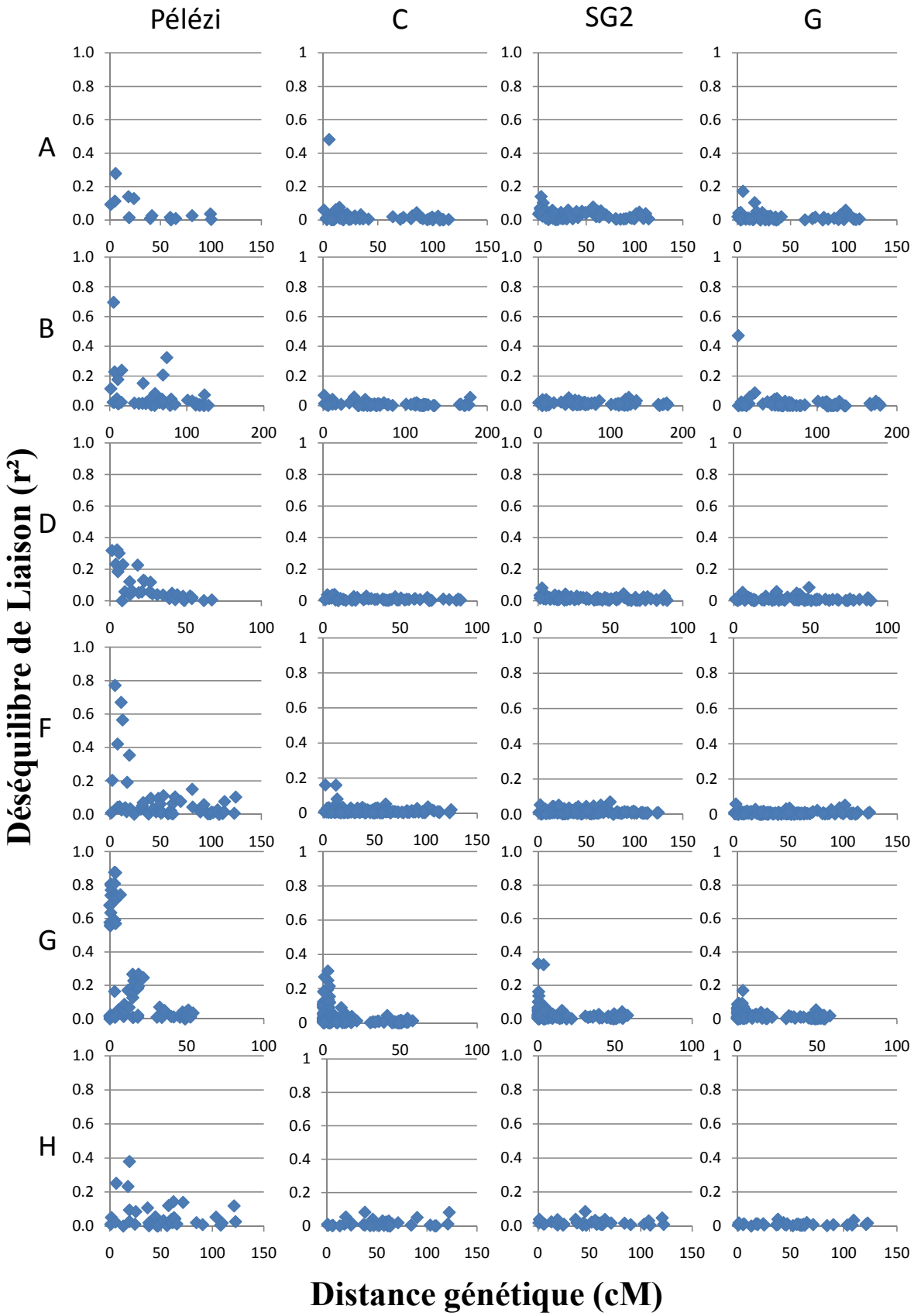


Figure 4.10 : Décroissance de r² en fonction de la distance génétique en centiMorgan pour 6 groupes de liaison pour quelques groupes de diversité.

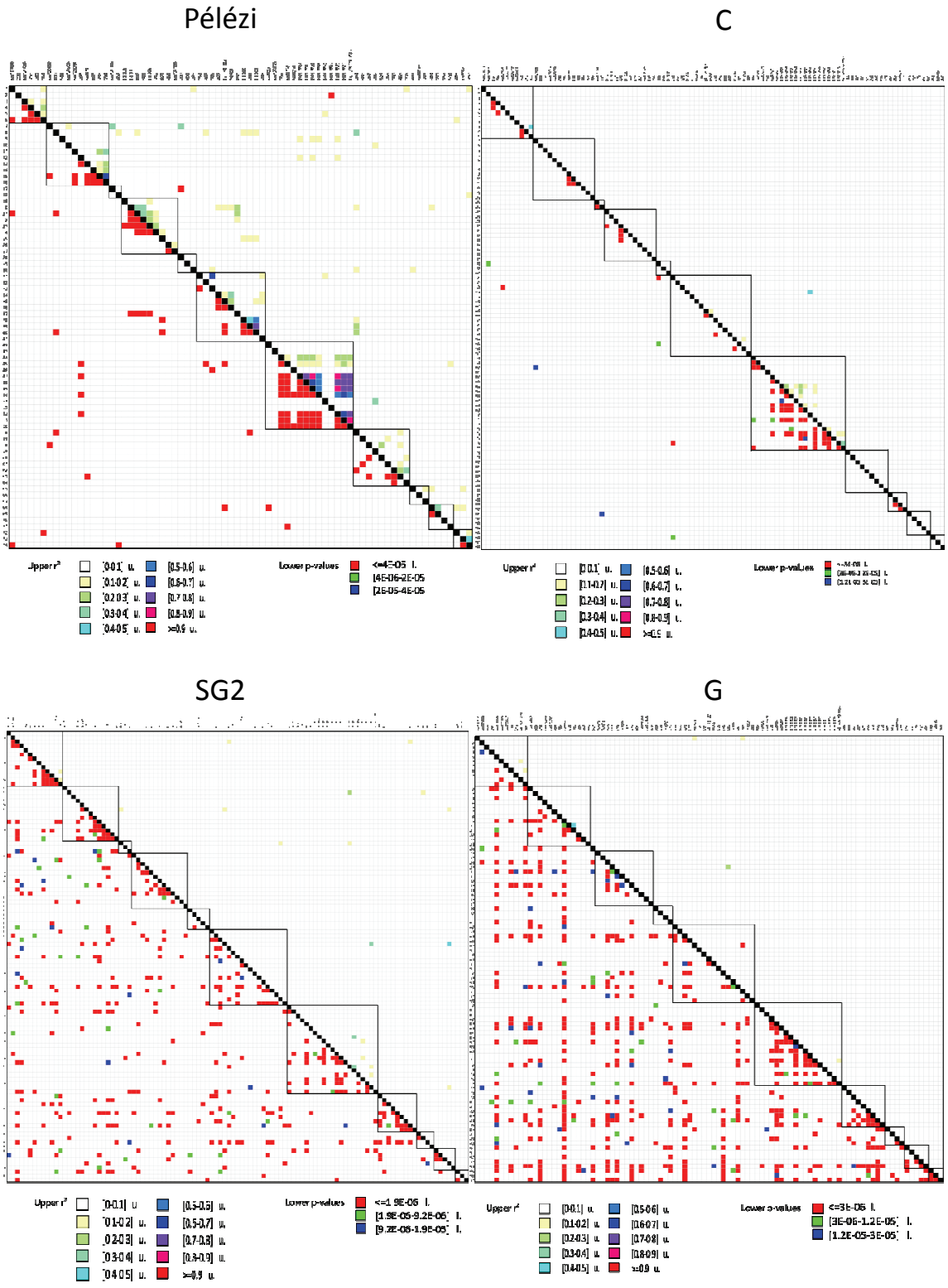


Figure 4.11 : Matrice de DL pour les groupes Pélézi, C, SG2 et G. Au-dessus des diagonales valeurs de r², en dessous des diagonales valeurs des p-values associées aux tests exacts. Les seuils de r² sont donnés en légende. Les seuils de p-values sont les seuils de 1, 5 et 10% après correction de Bonferroni. Les carrés mis en évidence dans la matrice correspondent aux différents groupes de liaison de la carte génétique

Discussion

Nous avons, dans ce chapitre, validé notre étude de structure réalisée au chapitre précédent avec un nombre beaucoup plus important de marqueurs liés et non liés. Les analyses de déséquilibre de liaison entre les différentes paires de marqueurs ont confirmé l'importance primordiale de la prise en compte de la structure dans les études d'association, celle-ci augmentant de manière drastique le nombre d'associations significatives à l'échelle du génome, y compris entre marqueurs appartenant à des groupes de liaison différents. Nous avons également mis en évidence la diversité de cas présente dans nos populations, à la fois en ce qui concerne la structure et l'intensité du déséquilibre de liaison. Nous avons enfin mis en évidence une stochasticité du déséquilibre de liaison à l'échelle du génome avec des différences marquées entre groupes de liaison.

La structure génétique chez C. canephora, comment la contrôler dans les études d'association ?

Nous avions avancé au chapitre précédent qu'un nombre de marqueurs non liés relativement faible, de l'ordre de 15 à 20, pouvait permettre d'analyser rapidement la structure générale de notre espèce et de différencier de manière claire les différents groupes de diversité mis en évidence, mais qu'un nombre de marqueurs plus important pourrait être nécessaire pour étudier la structure à un niveau inférieur.

Nous confirmons cette hypothèse et il apparaît, de part les résultats obtenus dans cette analyse, notamment sur la diversité des Guinéens, qu'un nombre important de marqueurs de « contrôle » est nécessaire pour arriver à séparer la structure fine en populations et évaluer les apparentements. Il semble que nous sommes dans un cas assez proche de celui observé pour le maïs où un set de 89 marqueurs microsatellites est utilisé pour étudier la structure et l'apparentement (Flint-Garcia *et al.*, 2005) sur 302 lignées.

L'importance de la prise en compte de ces 2 paramètres (structure et apparentement) dans les études d'association a été récemment soulignée, ces deux paramètres étant les principales causes de détection de fausses associations significatives. Il apparaît donc important de disposer de données substantielles sur des marqueurs à priori neutres afin de pouvoir inférer des matrices de structure et d'apparentements et de les incorporer aux modèles de génétique d'association utilisés.

Le déséquilibre de liaison chez Coffea canephora : une complexité insurmontable pour les études d'association?

Nous avons utilisé pour cette étude un déséquilibre de liaison haplotypique, basé sur des reconstructions d'haplotypes. Cette mesure à été choisie plutôt qu'un déséquilibre de liaison génotypique comme la mesure décrite par (Weir, 1996) et que nous avions précédemment utilisée (Cubry, 2005) pour le gain de puissance apporté (Barnaud *et al.*, 2006).

Les populations naturelles de *Coffea canephora* sont de taille relativement faible. De plus, de par le système de reproduction allogame strict de notre espèce, différents niveaux d'apparentement sont présents, créant des structures génétiques complexes à l'échelle même de la population. Cette structure « populationnelle » se superpose à celle plus importante mise en évidence au chapitre précédent, confirmée et précisée dans ce chapitre. L'ensemble de ces spécificités rendent l'étude du déséquilibre de liaison particulièrement complexe pour notre espèce. En effet le graphique représentant l'ensemble des valeurs de r^2 et *D'* à l'échelle des 356 génotypes montre bien l'importance de l'effet de la structure sur la détection d'associations entre des marqueurs non liés, rejetant la possibilité de travailler en étude d'association sur l'ensemble des génotypes en utilisant des modèles de corrélation simples qui ne prendraient pas en compte les effets de ce paramètre sur les associations détectées.

Ce résultat est confirmé par le grand nombre d'associations significatives entre marqueurs localisés sur des groupes de liaison différents pour les 356 génotypes par opposition aux résultats obtenus sur les groupes de diversité.

Afin d'étudier au mieux la dynamique du déséquilibre de liaison chez *C. canephora*, il semble donc nécessaire de se placer au niveau des populations. En effet, les déséquilibres les plus élevés entre marqueurs liés ont été obtenus sur les populations Pélézi, Nana (groupe C) et deux sous-groupes de Guinéens (Gsub1 et Gsub2) correspondant plus ou moins à des populations naturelles.

Nous avons pu mettre en évidence une forte variabilité du déséquilibre de liaison entre les différents groupes, avec une importante part de déséquilibre résiduel intergroupe de liaison, au sein des SG1 notamment, pouvant potentiellement mener à des faux positifs en étude d'associations, même à des faibles niveaux de structure génétique. L'importance de ce déséquilibre de liaison « génomique », par opposition à un déséquilibre de liaison local, est variable selon les groupes et nous pensons qu'une bonne prise en compte de la structure et de

l'apparentement dans les approches d'études d'association devrait permettre de le contrôler. Néanmoins en ce qui concerne le groupe SG2, les faibles valeurs de r^2 obtenues suggèrent qu'au-delà d'un possible effet de la structure, nous nous sommes placés à une échelle de distance trop importante pour pouvoir détecter des associations.

La détection de blocs de déséquilibre de liaison permet d'envisager l'utilisation de « marqueurs-étiquettes », c'est-à-dire de diminuer le nombre de marqueurs requis pour les études d'association au niveau génome entier en ne sélectionnant qu'un marqueur par bloc supposé de DL. Néanmoins cette approche nécessite une très bonne connaissance de la dynamique du DL et un nombre de marqueurs suffisamment important pour permettre de faire un choix parmi ceux-ci. A l'heure actuelle nos connaissances ne permettent pas de telles applications mais le développement des techniques de génotypage haut-débit permettent de les envisager à moyen terme. Ces approches pourraient permettre d'envisager sereinement des études « scan génome entier » y compris sur des échantillons montrant un DL relativement modéré.

Nous avons utilisé une correction de Bonferroni afin de ne considérer que les valeurs réellement significatives. Néanmoins cette correction est très conservatrice et peut mener à une perte de puissance importante dans les études d'association. De nombreuses autres corrections ont été avancées au cours des dernières années dans la littérature, cependant aucune ne semble réellement satisfaisante. Nous avons par ailleurs montré que dans notre cas cette correction permet principalement d'éliminer un certain nombre de valeurs de déséquilibre entre marqueurs non liés ou des valeurs très faibles. La plupart du temps dans les études d'association nous n'aurons pas à réaliser cette correction puisque la source principale d'erreur (la structure génétique) sera contrôlée. Une réflexion plus poussée sur ces questions ainsi qu'une veille bibliographique sur les nouveaux modèles proposés sera nécessaire. Il semble que les modèles prenant en compte à la fois la structure et l'apparentement dans les études d'association soit une avancée majeure dans ces démarches, permettant d'augmenter à la fois la puissance et la résolution de telles études.

Quelles populations cibles et quelles approches ?

Notre travail avait pour but une première évaluation du déséquilibre de liaison au niveau pan-génomique de *C. canephora*. Nous avons ainsi pu mettre en évidence une variation très importante du DL entre populations. Il semble que les populations naturelles comme Pélézi ou les caféiers de la Nana (groupe C) montrent une étendue de DL moyenne à

élevée, de l'ordre de 5 à 25 cM. Dans ces populations il semble envisageable de mener des études de type « scan génome entier ». Néanmoins la stochasticité du DL entre groupes de liaison ainsi que sa sensibilité aux faibles fréquences alléliques nous mène a émettre quelques réserves sur ce type d'approche. Une première évaluation de ces techniques pourra être faite en comparant les associations détectées à celles déjà mise en évidence par des approches de cartographie, notamment sur des caractères faiblement polygéniques dans un premier temps, comme par exemple la caféine ou la granulométrie.

Les populations de type amélioré telles que SG2 semblent avoir subi un brassage génétique relativement important avec une plus grande diversité que les populations naturelles (voir chapitre 3) ainsi qu'un DL quasiment indétectable à l'échelle où nous nous sommes situés. Par conséquent ces populations semblent plus adaptées à des approches de type régions candidates ou gènes candidats.

Conclusion

Nous avons présenté ici les premiers résultats d'analyse de DL au niveau génome entier sur *C. canephora*. Comme chez de nombreuses espèces, le DL observé est très variable à la fois entre populations et entre régions différentes du génome. Nous avons pu mettre en évidence des populations potentiellement utilisables pour les 2 grandes approches de la génétique d'associations, « scan génome entier » ou « région candidate ou gène candidat ». La structure génétique des populations chez *C. canephora* apparaît comme complexe et nécessite un nombre assez important de marqueurs de contrôle permettant d'inférer à la fois des matrices de structure et des matrices d'apparentement pour les incorporer aux modèles statistiques de détection d'associations. Il semble possible de mener des études d'association sur notre espèce en étant relativement confiant et en les intégrant de manière efficace aux schémas de sélection préexistants.

Chapitre 5 : Le déséquilibre de liaison à l'échelle physique par microsatellites et polymorphismes de séquences

Introduction

Avec le développement rapide de nouvelles techniques de séquençage et de génotypage, la possibilité d'études de déséquilibre de liaison et d'études d'association basées sur des polymorphismes de séquences (par exemple par génotypage de SNPs) est envisageable. Afin d'évaluer la possibilité de telles approches, nous avons étudié à l'aide de quelques fragments de séquences et de microsatellites une zone précise du génome. Cette étude, basée sur 3 fragments de séquences choisis au hasard dans le clone BAC 111O18 et 5 fragments d'un gène de Saccharose Synthéthase (SUSY), permet une première approche ainsi qu'une comparaison entre les propriétés des SNPs et des microsatellites pour les études de DL et d'associations. De plus elle nous donne également une idée de la décroissance du DL avec la distance physique en paire de bases et non plus en centiMorgan.

Matériel et Méthodes

Matériel végétal

Nous avons choisit un échantillon de 2 groupes de diversité pour mener cette étude. Au total 23 génotypes du groupe C et 25 génotypes du groupe G au sens large (incluant 2 génotypes de la population Pélézi) ont été retenus. Le choix de ces génotypes a été guidé par l'utilisation d'une procédure de ré-échantillonnage d'arbre implémentée dans le logiciel DARwin. Deux méthodes peuvent être utilisées dans ce logiciel afin de réaliser des ré-échantillonnages en vu d'études de déséquilibre de liaison. La première de ces méthodes utilise un set de marqueurs non liés afin de minimiser le DL global en fonction de la taille d'échantillon retenue, en comparant cet échantillonnage à un échantillonnage au hasard. Cette méthode n'est pas adaptée pour nous puisque nous avons un grand nombre de marqueurs liés dans notre étude. La seconde méthode va essayer d'éliminer au maximum les redondances entre génotypes, permettant d'obtenir l'arbre de diversité possédant les plus grandes

longueurs de branches possibles en fonction de la taille d'échantillon choisie. Le choix de la taille est totalement arbitraire et dépend donc de l'organisation du travail et des connaissances préliminaires que l'on peut avoir sur nos génotypes (Perrier, communication personnelle). Cette méthode est dite de « max-length sub tree ».

Le ré-échantillonnage par la méthode de « max-length sub tree » pour le choix des 48 génotypes a été réalisé sur l'étude de diversité présentée au chapitre 3, de manière séparée pour les groupes C et G. Pour le groupe G nous avons choisi de mettre les génotypes Guinéens cultivés en « forcé » car ce groupe présentait une diversité intéressante couvrant quasiment l'ensemble des populations naturelles (voir chapitre 3).

Validation de l'échantillonnage

Afin de valider l'échantillonnage réalisé, des arbres Neighbor-Joining basés sur les dissimilarités calculées entre individus à l'aide des marqueurs polymorphes parmi les 108 marqueurs de l'étude précédente ont été réalisés à l'aide de DARwin. La forme de l'arbre a été analysée afin d'évaluer la présence de structure. Nous avons également étudié le déséquilibre de liaison (r^2) au niveau pan-génomique avec les mêmes marqueurs.

Etude du déséquilibre de liaison au sein du clone BAC 111O18 et du gène Susy2

Microsatellites du clone BAC 111O18

Les microsatellites utilisés dans cette étude sont ceux décrits au chapitre précédent développés à partir de la séquence consensus du clone BAC 111O18. Les distances entre marqueurs utilisées ici sont des distances physiques en paires de base en contraste avec les distances génétiques approximatives utilisées pour le chapitre précédent.

Fragments séquencés

Choix

Trois fragments ont été choisis au hasard sur l'ensemble de la séquence consensus du clone BAC 111O18. Ces fragments se situent au début (1275pb-2015pb), en milieu (89018pb-89749pb) et en fin (169892pb-170600pb) de la séquence consensus de ce clone. 5 fragments du gène Susy2 (Pot, communication personnelle) ont été utilisés pour étudier le DL au sein de ce gène. Ces fragments contiennent des introns et des exons et sont répartis sur toute la longueur du gène. Ce gène a été analysé précédemment et ne semble pas contraint par la

sélection (Bouchet, 2005). Le DL détecté dans cette zone sera donc à priori assez informatif sur l'étendue et l'intensité du DL mesurable sur de faibles distances physiques.

Les produits PCR obtenus ont été séquencés dans le sens forward par la société GATC.

Traitement des séquences

Les séquences ont été nettoyées et alignées à l'aide du logiciel CodonCode Aligner® 1.6.3. La recherche de polymorphismes SNP et indel a été réalisée manuellement par étude des chromatogrammes à l'aide de ce même logiciel.

Reconstruction de phase

La reconstruction de phase a été réalisée de la même manière qu'au chapitre précédent en utilisant le logiciel PHASE et l'ensemble des 48 génotypes. Cinq runs indépendants du programme ont été réalisés et le run ayant obtenu la meilleure vraisemblance a été conservé pour la suite des analyses.

Analyse du déséquilibre de liaison

Les analyses statistiques de déséquilibre de liaison ont été réalisées suivant la même méthode que celle présentée au chapitre 4. Dans ces analyses, nous avons choisi de considérer de manière séparée les résultats obtenus par microsatellites et par polymorphismes de séquence. Ces deux types de marqueur ayant des propriétés et des mécanismes d'évolution différents sont difficilement comparables sur les mêmes analyses, pouvant mener par exemple à des détections de blocs de DL uniquement dus à la présence contigüe de plusieurs marqueurs du même type. Nous avons néanmoins conservé l'ensemble des marqueurs polymorphes dans les représentations matricielles du DL afin de pourvoir avoir une vision d'ensemble du DL à cette échelle à l'aide des deux types de marqueur possibles.

Résultats

Echantillonnage

23 génotypes du groupe C et 25 génotypes du groupe G ont été retenus pour cette étude après ré-échantillonnage par la méthode de « max-length sub tree » implémentée dans DARwin 5.0. La liste de ces génotypes est donnée en Tableau 5.1.

Tableau 5.1 : Liste des 48 génotypes choisis par rééchantillonage pour minimiser la structure au sein des groupes de diversité C et G

Génotypes C	Génotypes G
c4001	g1003
c4002	g1032
c4003	g2014
c4004	g2020
c4005	g3001
c4009	g3002
c4011	g3004
c4012	g3005
c4013	g3007
c4014	g3008
c4018	g3009
c4020	g3010
c4021	g3011
c4024	g3012
c4025	g3013
c4026	g3017
c4027	g3018
c4030	g3019
c4031	g3020
c4033	g3021
c4036	g4002
c4037	g5012
c4038	g5017
	g5019
	g6019

Structure et diversité des 48 individus et des 2 groupes

Arbre de Neighbour-Joining et diversité pour les 48 génotypes

L'arbre de Neighbour-Joining généré pour les 48 génotypes à l'aide des 107 marqueurs polymorphes parmi les 108 de l'étude précédente est donné en Figure 5.1. On peut constater sur cette figure une très nette distinction entre les groupes C et G. Les distances génétiques semblent plus importantes au sein du groupe G qu'au sein du groupe C. Ceci peut s'expliquer par le fait que l'origine des génotypes représentant le groupe C dans notre travail est une seule population alors que le groupe G est composé de plusieurs populations naturelles et d'une population composite de génotypes prospectés en plantation.

Ce résultat est cohérent avec la structure génétique globale de l'espèce décrite précédemment.

Un nombre moyen d'allèles par marqueur de 6,48 a été obtenu pour cet échantillon, ce qui représente environ la moitié de la valeur obtenue pour l'espèce sur 39 marqueurs (chapitre 3). L'hétérozygotie attendue moyenne est de 0,61, une valeur moyenne en regard de l'espèce, qui semble indiquer une diversité assez importante dans l'échantillon retenu.

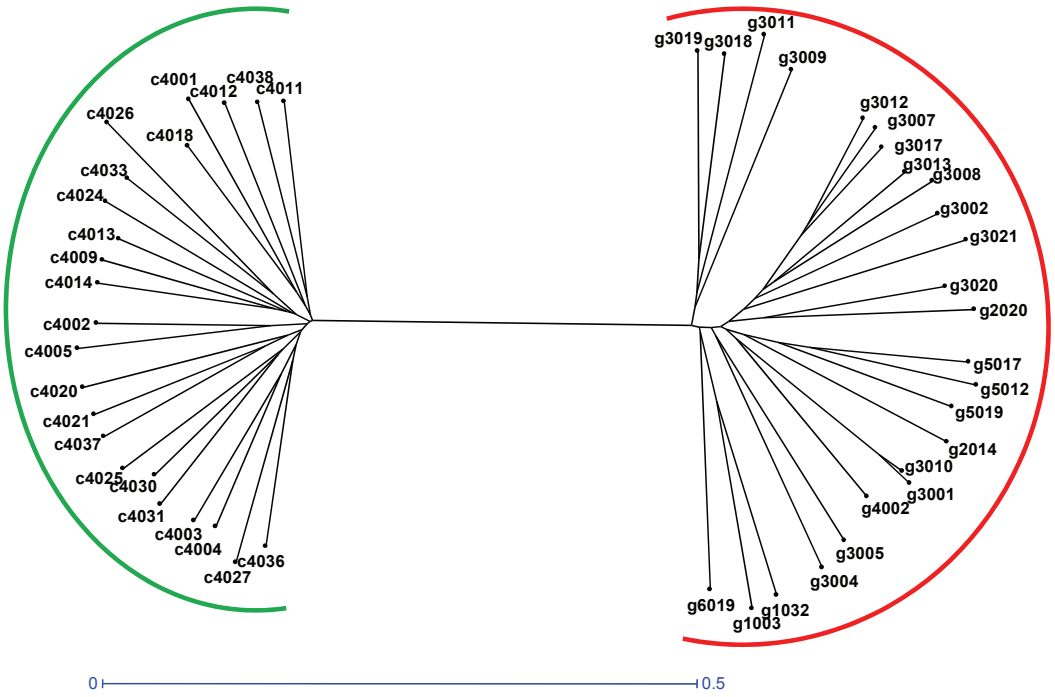


Figure 5.1 : Arbre NJ des 48 individus utilisés dans cette étude, à gauche les génotypes appartenant au groupe C (en vert), à droite ceux appartenant au groupe G (en rouge). L'échelle de distance donnée correspond à l'index de Simple-Matching utilisé pour calculer les dissimilarités entre individus à l'aide du logiciel DARwin (Perrier *et al.*, 2003)

Arbre de Neighbour-Joining et diversité pour les groupes C (23 individus) et G (25 individus)

Les arbres correspondants à ce paragraphe sont donnés en Figure 5.2. Pour le groupe C, on constate une répartition des individus en une étoile quasi-parfaite, validant le sous-échantillonage réalisé au sein du groupe C. Le nombre moyen d'allèles par marqueur est de 3,89, ce qui est légèrement inférieur à celui obtenu sur un échantillon de 43 individus de cette

population (4,74). En revanche l'hétérozygotie attendue est quasiment identique (0,5 dans cet échantillon contre 0,49 dans le chapitre 3), ce qui montre qu'une part importante de la diversité de cette population a été capturée dans notre échantillon, avec sans doute une baisse du nombre moyen d'allèles due à une perte d'allèles rares. Nous pouvons donc valider *a posteriori* notre sous-échantillon comme un échantillon représentatif peu structuré de la population originale.

Pour le groupe G, la répartition des individus selon l'arbre NJ est globalement satisfaisante avec néanmoins quelques redondances dues au fait que nous avons forcé l'ensemble des génotypes prospectés en plantation qui semblaient former une sorte de continuum sur la diversité de certaines populations guinéennes (voir le chapitre 3). Le nombre moyen d'allèles obtenu est de 5,2 contre 7,4 pour l'ensemble du groupe G obtenu au chapitre 3. L'hétérozygotie attendue est quant à elle de 0,5, très comparable à la valeur de 0,53 obtenue au chapitre 3. La même conclusion pour ce sous-échantillon que pour celui du groupe C peut être déduite, avec vraisemblablement une baisse du nombre moyen d'allèle principalement due à une perte d'allèles rares mais une diversité globale représentative du groupe G.

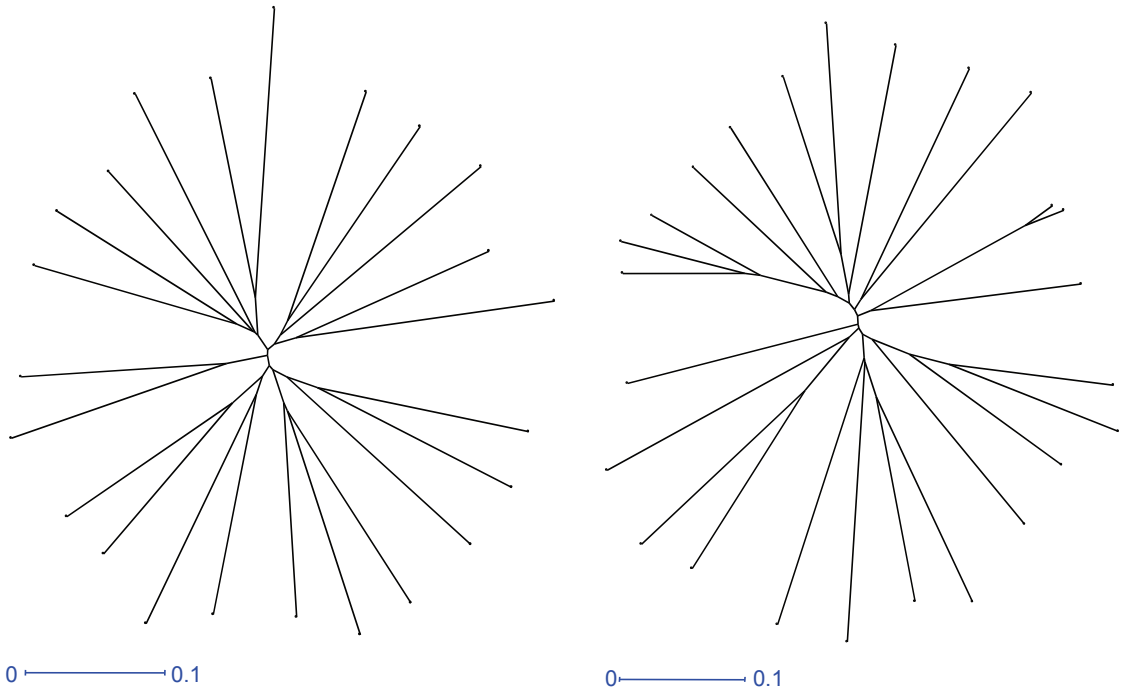


Figure 5.2 : Arbres NJ des sous-échantillons des groupes C (à gauche) et G (à droite). Les échelles de distance données correspondent à l'index de Simple-Matching

Déséquilibre de liaison au niveau pan-génomique

Afin de valider définitivement notre ré-échantillonnage nous avons calculé le déséquilibre de liaison pangénomique de la même manière qu'au chapitre précédent. Les résultats de cette analyse sont présentés en Figure 5.3.

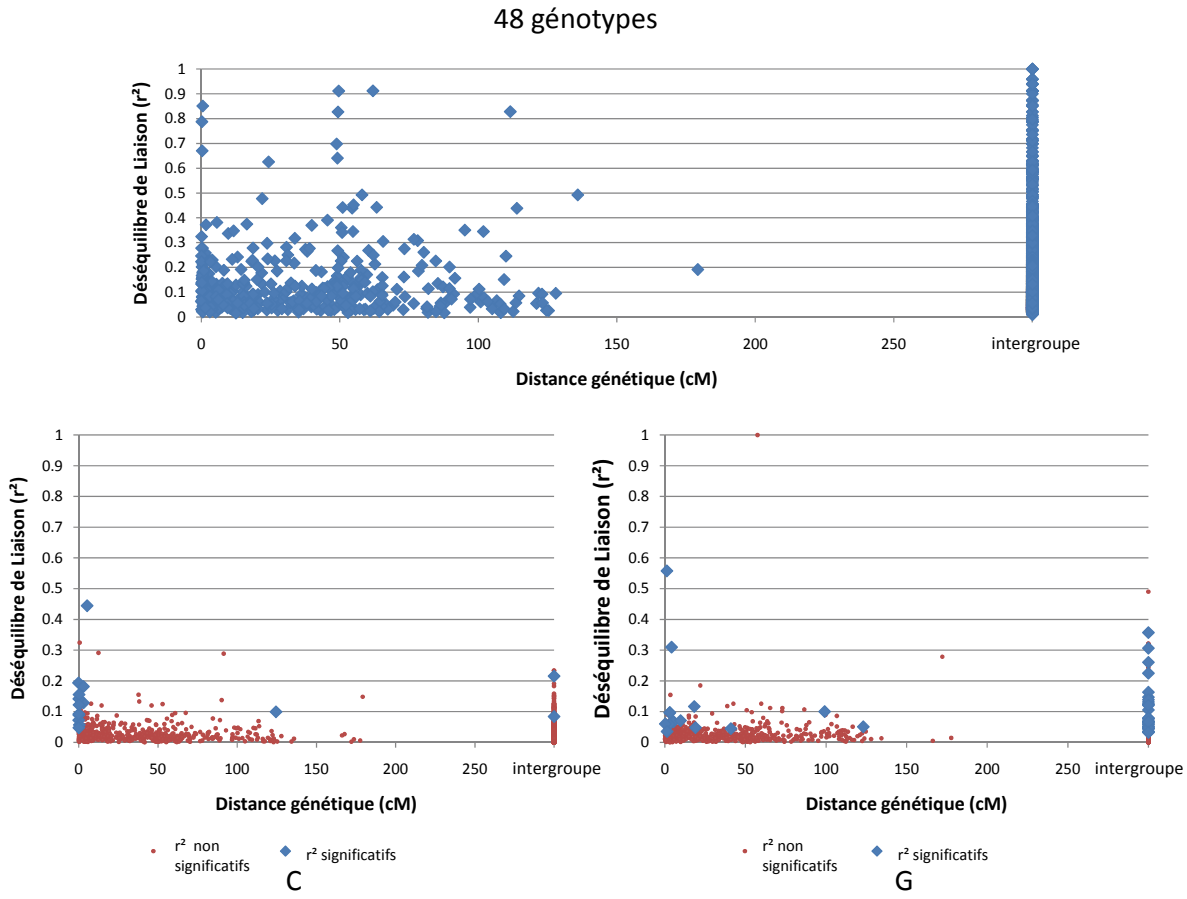


Figure 5.3 : r² en fonction de la distance génétique en centiMorgan pour les 48 génotypes et pour les 2 groupes. En bleu les valeurs significatives au seuil de 5% après correction de Bonferroni, en rouge les valeurs non significatives. Pour les 48 génotypes, seules les valeurs significatives ont été représentées

Ces résultats confirment ceux présentés au chapitre précédent, avec un très fort effet de la structure sur les déséquilibres détectés. Quand on s'intéresse aux 2 sous-échantillons, on retrouve la quasi absence d'associations significatives entre marqueurs non liés pour le groupe C alors qu'un plus grand nombre est détecté pour le groupe G. On constate globalement le faible nombre d'associations significatives pour les 2 groupes, avec majoritairement des associations entre marqueurs très proches. Il semble que la réduction du nombre de génotypes dans les études de DL ait une influence directe et très importante sur la puissance de détection des associations.

Diversité et polymorphismes des marqueurs dans le clone BAC 111O18

Un total de 15 marqueurs microsatellites et de 40 marqueurs de polymorphismes de séquences a été mis en évidence. La Figure 5.4 présente la position de ces marqueurs le long du clone BAC. L'ensemble de ces marqueurs se sont révélés polymorphes sur les 48 individus. Au sein des 2 sous-échantillons en revanche un certain nombre de ces marqueurs était fixé. Nous avons donc obtenu 13 marqueurs microsatellites et 22 marqueurs de séquence polymorphes pour le groupe C, 12 microsatellites et 25 marqueurs de séquence pour le groupe G. Les indices de diversité calculés pour les 48 individus et les 2 sous échantillons, incluant le nombre d'allèles par marqueurs, l'hétérozygotie attendue et observée sont donnés en Annexe A.5.1.

DL par microsatellites au sein du clone BAC 111O18

Le déséquilibre de liaison entre marqueurs microsatellites au sein du clone BAC 111O18 est présenté en Figure 5.5.

Les valeurs de r^2 calculées sont faibles à moyennes. Des associations significatives peuvent être détectées sur toute la longueur du clone (180kb environ). Les valeurs de r^2 pour l'ensemble des 48 génotypes sont beaucoup plus élevées que celles détectées au sein des 2 sous-échantillons. Ce clone représente environ 0,3 cM (1cM est environ égal à 570 kb) et il est donc cohérent de trouver des valeurs d'associations significatives sur la totalité du clone en considérant les résultats du chapitre précédent. Les faibles valeurs observées doivent être pondérées par les faibles effectifs des populations de cette étude, les effets de ces effectifs au niveau du clone devant être les mêmes qu'au niveau pan-génomique, c'est-à-dire une baisse du nombre d'associations détectées due à une perte de puissance.

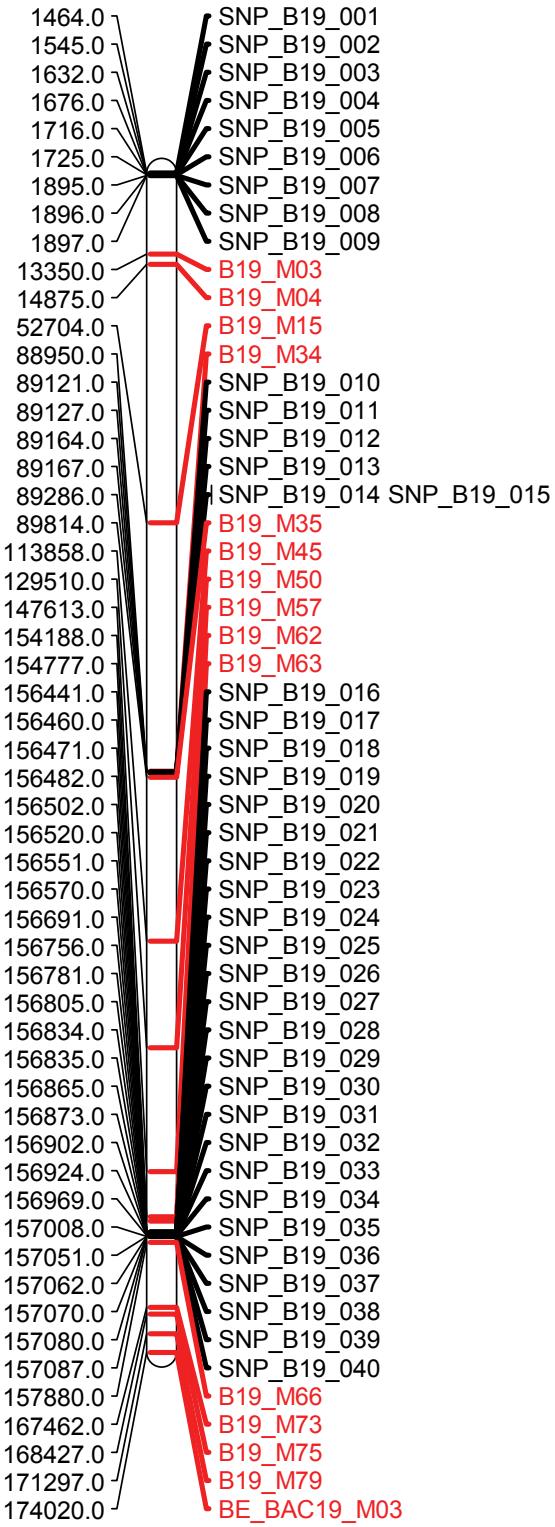


Figure 5.4 : Positionnement des marqueurs mis en évidence dans le clone BAC 111018. Les positions indiquées sont les positions en paires de base sur la séquence consensus du clone BAC. Les marqueurs en noir sont des polymorphismes de séquence, les marqueurs en rouge sont des microsatellites

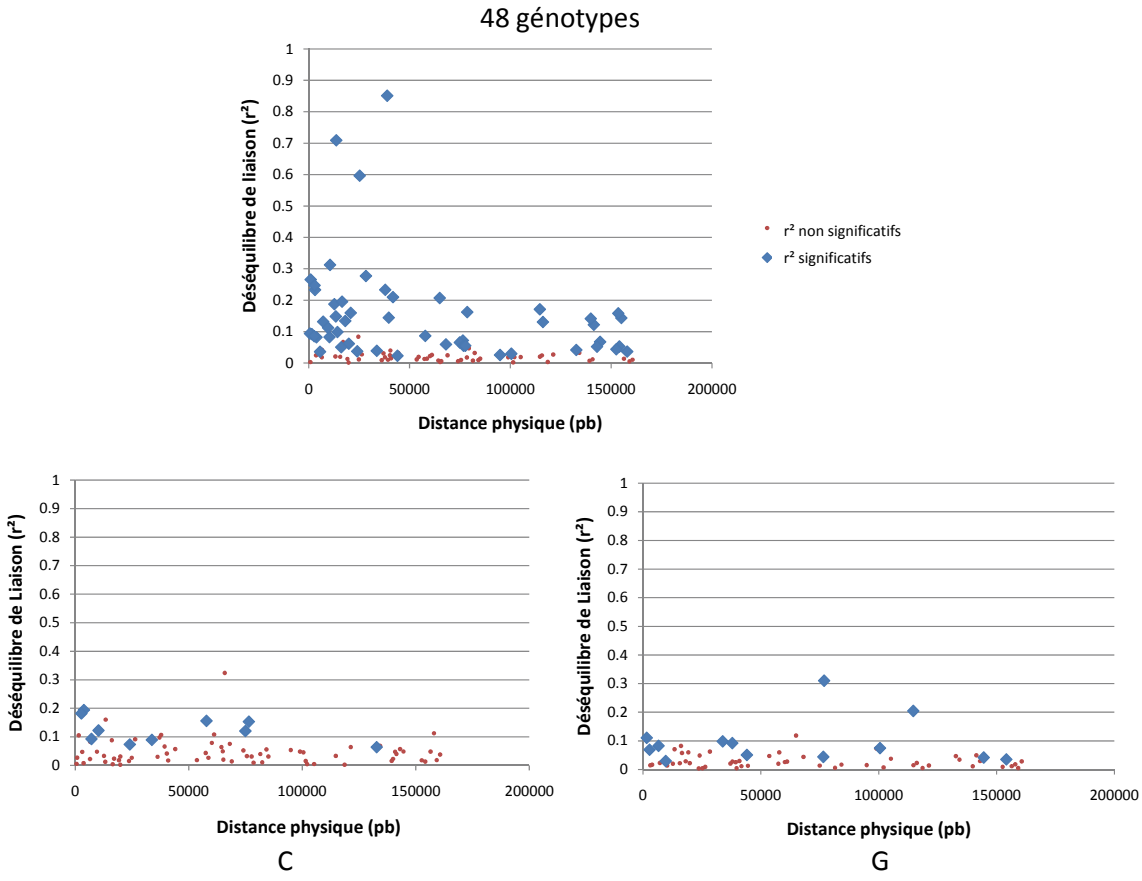


Figure 5.5 : Déséquilibre de liaison (r^2) entre locus microsatellites en fonction de la distance physique en paire de bases (pb) au sein du clone BAC 111O18. En bleu les valeurs significatives au seuil de 5% après correction de Bonferroni, en rouge les valeurs non significatives.

DL par séquences

Les résultats de l'analyse des déséquilibres de liaison entre polymorphismes de séquences sont présentés en Figure 5.6. Dans cette analyse, on n'observe pas de déséquilibres significatifs pour les 2 sous-échantillons au-delà de 100kb. En revanche, pour les 48 génotypes, des associations significatives sont détectées tout au long du clone. Néanmoins le graphique présenté pour les 48 génotypes nous laisse penser que les associations détectées au-delà des 100kb sont principalement dues à la structure (augmentation des valeurs de r^2 et du nombre d'associations significatives non cohérentes avec la décroissance observée du DL sur les 100 premiers kb). Si le DL entre polymorphismes de séquences semble donc être significatif à moins longue distance que les DL observés par microsatellites, les valeurs de r^2

sont elles beaucoup plus élevées. Ceci est à nuancer à cause de la fragmentation des données en seulement 3 fragments séquencés sur l'ensemble du clone.

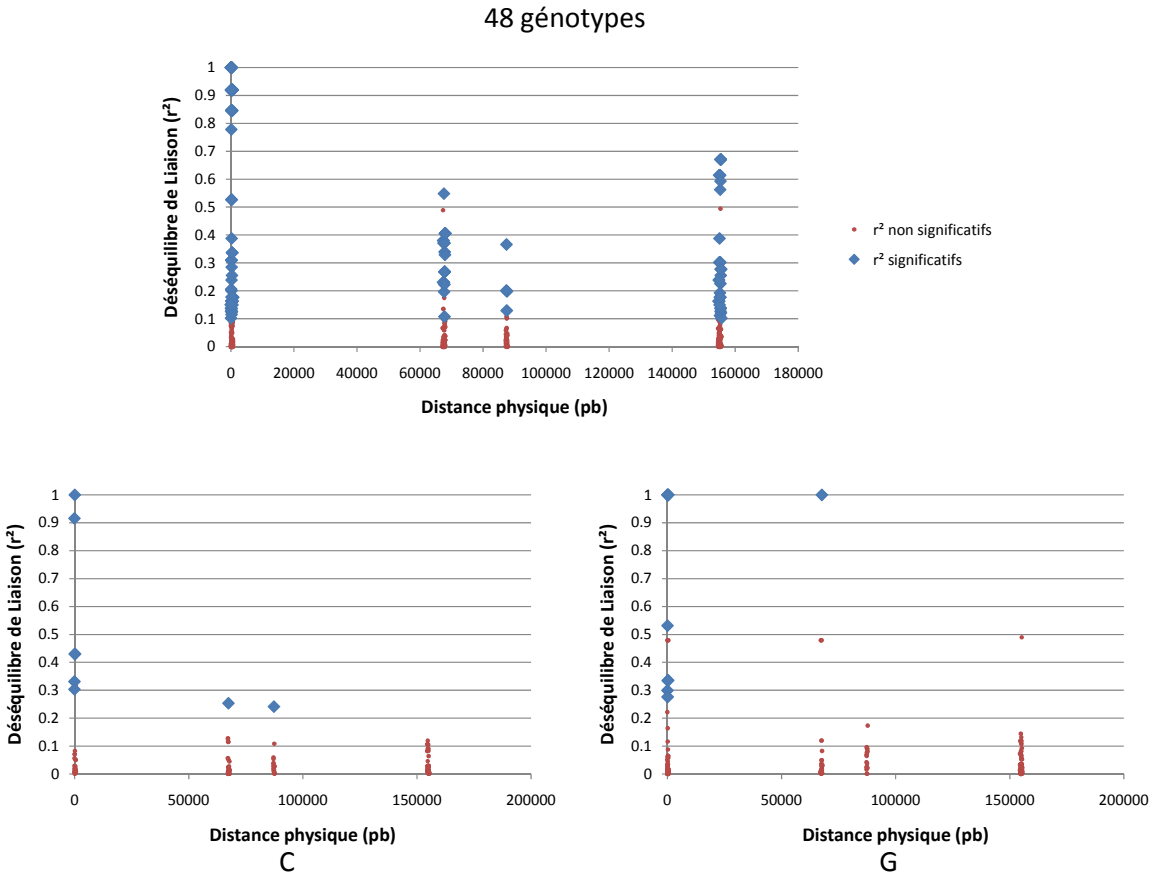


Figure 5.6 : Déséquilibre de liaison (r^2) entre polymorphismes de séquence en fonction de la distance physique en paire de bases (pb). En bleu les valeurs significatives au seuil de 5% après correction de Bonferroni, en rouge les valeurs non significatives

Comparaison DL séquences/microsatellites

Les profils de DL par microsatellites et par séquences montrent donc des différences importantes. Si la significativité des associations baisse plus rapidement avec les polymorphismes de séquences qu'avec les microsatellites, les valeurs de r^2 sont beaucoup plus faibles avec ces seconds marqueurs. Ceci est principalement dû au nombre d'allèles plus important dans le cas des microsatellites et donc à la plus grande efficacité de la recombinaison entre ces marqueurs. En effet, on peut aisément imaginer que si le nombre d'allèle est plus important, le nombre de recombinaisons efficaces qui seront détectées et qui pourront diminuer le DL entre deux marqueurs sera plus important.

La Figure 5.7 présente les matrices de DL au niveau du clone Bac 111O18.

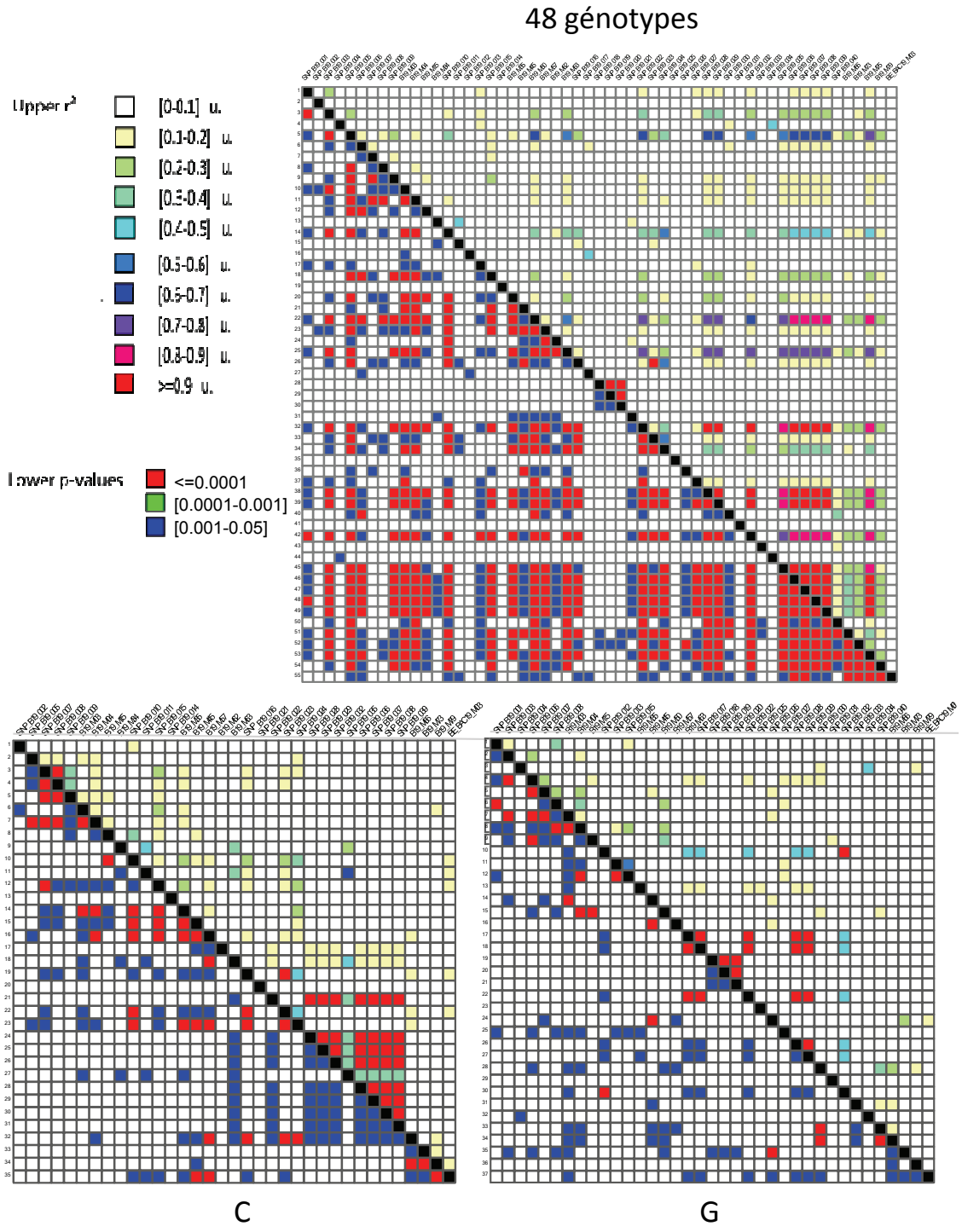


Figure 5.7 : Matrices de DL du clone BAC 111O18 entre marqueurs 2 à 2. Au-dessus des diagonales sont présentées les valeurs de r² (les seuils sont donnés en légende), en dessous les valeurs des p-values associées aux tests exacts de Fisher aux seuils de 1 pour 1000, 1 pour 100 et 5 pour 100 (sans correction)

Ces matrices confortent les résultats précédemment évoqués. Le grand nombre d'associations significatives pour les 48 génotypes est bien dû principalement à la structure. Au niveau des 2 sous-échantillons la majorité des associations à faibles p-values se retrouvent proches de la diagonale. On observe 2 grands blocs de DL pour le groupe C. Pour le groupe G, ces blocs semblent moins évidents et plus restreints. Les valeurs de r^2 les plus importantes se retrouvent entre les polymorphismes de séquences. Les 2 groupes présentent donc des résultats très contrastés, validant les observations faites au chapitre précédent.

Polymorphismes et DL dans le gène Susy2

La position des polymorphismes de séquences observés le long du gène Susy2 est donnée en figure 5.8.

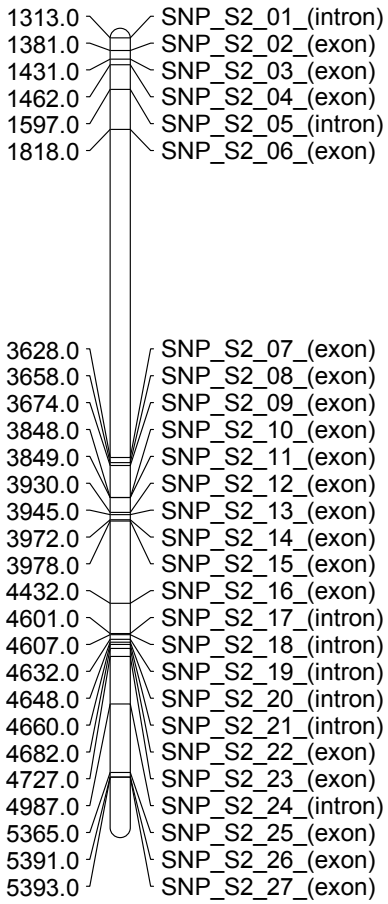


Figure 5.8 : Positionnement des polymorphismes mis en évidence dans le gène Susy2. Les positions en paires de bases sont données par rapport à la séquence consensus (Pot, communication personnelle). La position des polymorphismes dans les introns ou les exons est précisée

Les polymorphismes observés couvrent la quasi-totalité du gène et se trouvent indifféremment dans des introns ou des exons. Un certain nombre de ces polymorphismes sont fixés dans l'une ou l'autre des populations voire dans les deux, ce qui porte à 13 et 15 le nombre de marqueurs polymorphes pour les groupes C et G respectivement. Les graphiques des valeurs de r^2 en fonction de la distance sont donnés en Figure 5.9, les matrices correspondantes en Figure 5.10.

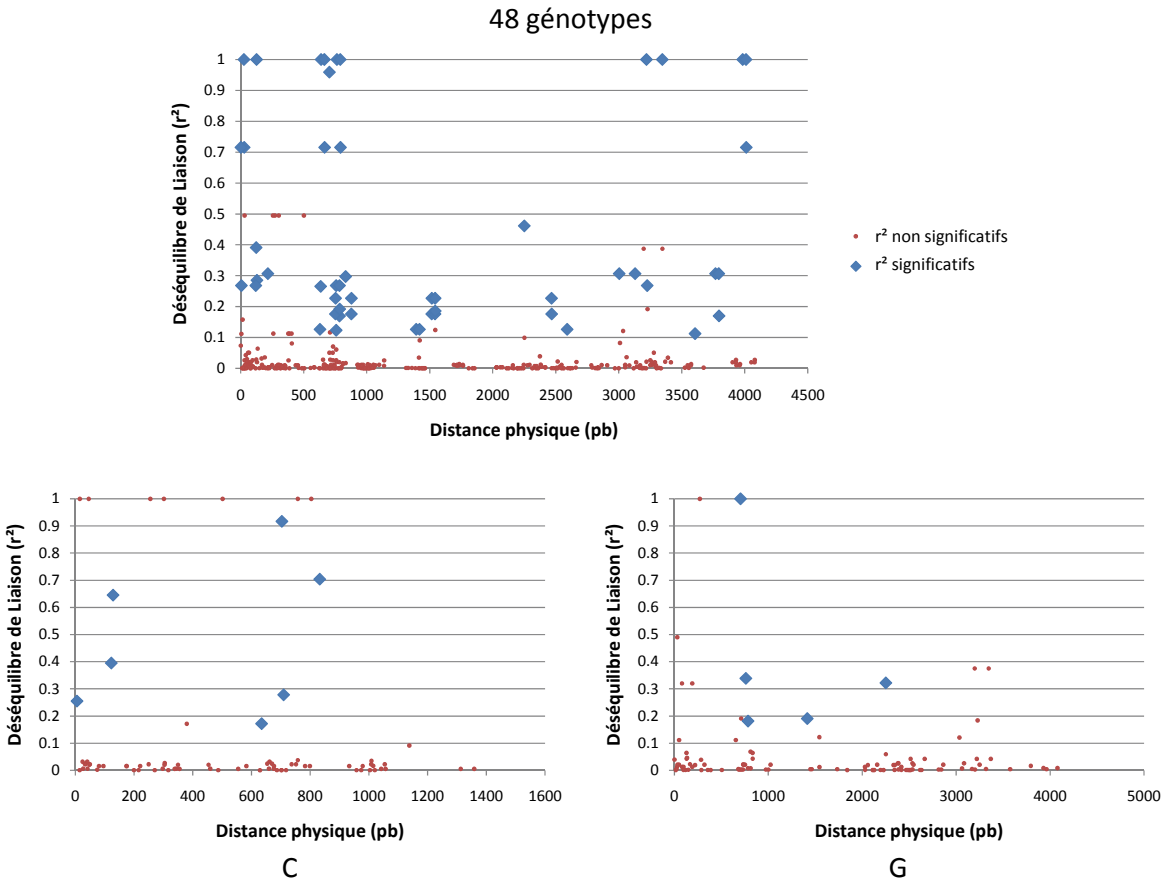


Figure 5.9 : Déséquilibre de liaison (r^2) entre polymorphismes de séquence en fonction de la distance physique en paire de bases

Sur l'ensemble des 48 génotypes des valeurs significatives de DL sont observées tout le long du gène. Pour les 2 groupes nous n'en observons que jusqu'à 800 paires de bases (groupe C) et 2440 paires de bases (groupe G). Ces derniers résultats sont à nuancer en constatant l'absence de couples de marqueurs au-delà de 1400 paires de bases pour le groupe C et la moindre importance du DL au niveau du groupe G (voir matrices de la Figure 5.10). On ne peut donc pas émettre de conclusions nettes sur le DL au niveau de ce gène avec notre

étude, des études complémentaires de séquençage sur un plus grand nombre d'individus seraient nécessaires pour préciser l'étendue et la dynamique du DL au sein du gène Susy2. Nous observons néanmoins encore une fois une grande différence de DL en fonction des populations et la sensibilité de nos études à la présence de structure génétique, avec un nombre de valeurs significatives beaucoup plus importante pour les 48 génotypes que pour les 2 groupes séparément, et ce malgré une correction de Bonferroni sur le seuil de signification des p-values associées au test exact.

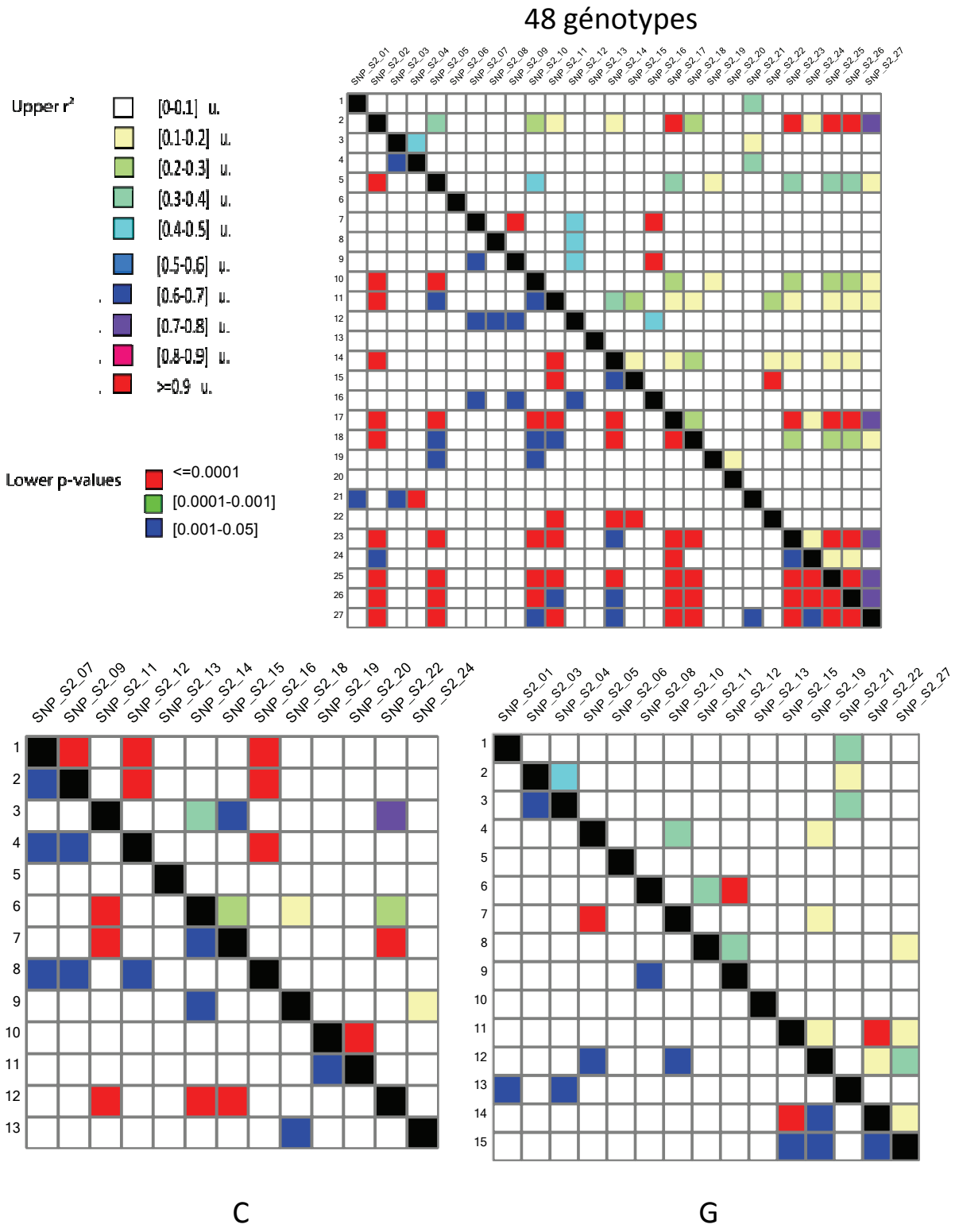


Figure 5.10 : Matrices de DL du gène Susy2. Au-dessus des diagonales sont présentées les valeurs de r², en dessous les valeurs des p-values associées au test exact de Fisher, les seuils de 1 pour 1000, 1 pour 100 et 5 pour 100 sans correction sont considérés

Discussion

Méthode d'échantillonnage

Nous avons validé la méthode de « max-length sub-tree » pour l'obtention d'échantillons les moins structurés possibles. Les arbres obtenus montrent en effet une structure moins prononcée que les arbres d'origine. Nous avons également montré la diminution du nombre moyen d'allèles dans les échantillons générés. Ceci est facilement explicable par la méthode d'échantillonnage utilisée qui ne s'intéresse qu'à la structure de l'arbre, par une approche que l'on peut qualifier de topographique, sans s'intéresser à la diversité en tant que telle. Cette méthode d'échantillonnage se démarque donc de méthodes de construction de core-collections telles que la stratégie M (Gouesnard *et al.*, 2001) qui ont pour principal but de maximiser la diversité en essayant de conserver le maximum d'allèles et augmentent artificiellement la fréquence des allèles rares. Néanmoins ces méthodes présentent toutes les deux des avantages et des inconvénients pour la définition de populations d'association. C'est pourquoi une réflexion sur l'intérêt et la mise en place de telles méthodes, tout en considérant les avancées méthodologiques permettant la prise en compte de la structure génétique dans les calculs d'associations entre marqueurs et caractères d'intérêt devra être menée. Les core-collections présentent par exemple des avantages en conservation des ressources génétiques et pourraient être utilisées pour différents objectifs.

L'intérêt des polymorphismes de séquence et des microsatellites pour les études de DL ou d'association sur C. canephora

Nous avons présenté dans cette partie une évaluation du déséquilibre de liaison basée sur des polymorphismes de séquences à 2 échelles, celle de quelques gènes (180kb) et celle d'un gène (5kb). Ces polymorphismes montrent globalement des valeurs de r^2 beaucoup plus élevées que celles trouvées à l'aide des microsatellites, en revanche les associations semblent être significatives à de plus courtes distances. La baisse plus rapide du DL entre polymorphismes de séquences par rapport à celle observée entre microsatellites est un résultat attendu pour des populations ayant eu une longue histoire de recombinaison. Les taux de mutation des polymorphismes de séquence étant très inférieurs à ceux des microsatellites, on s'attend à ce que les recombinaisons aient « brisé » le DL initial qu'il pouvait exister entre locus au niveau des séquences, celles-ci ayant potentiellement subi plus de recombinaisons (Remington *et al.*, 2001). Il semble donc d'après ces résultats que les polymorphismes de type

SNP serait plus performants pour des études de types « gène candidat », à de faibles distances physiques, alors que les microsatellites seraient plus adaptés pour des distances plus importantes, voire des approches génome entier. L'importance de la différence des valeurs de r^2 obtenue entre les SNP et les microsatellites peut-être expliquée par la sensibilité de r^2 aux allèles rares et par les faibles effectifs que nous avons utilisés pour cette étude. Il semblerait donc que pour une utilisation correcte des marqueurs microsatellites, il faille augmenter de manière préférentielle la taille des échantillons par rapport au nombre de marqueurs. Ce résultat est vérifié par la perte de puissance constatée entre les sous-échantillons des groupes C et G utilisés ici par rapport aux résultats obtenus dans le chapitre précédent. Il semblerait donc opportun d'établir des seuils de valeurs de r^2 différentes spécifiques à chaque type de marqueurs en fonction des résultats que nous venons d'exposer.

Les deux types de marqueurs sont potentiellement intéressants pour la génétique d'association chez *C. canephora* et possèdent des caractéristiques différentes qui demandent des approches adaptées. La disponibilité de marqueurs microsatellites en grand nombre et déjà cartographiés sur notre espèce permet d'espérer un développement rapide des études d'association ayant recours à ces marqueurs sur des populations ciblées. Par ailleurs les ressources génomiques disponibles permettent également d'envisager ce type d'approche à l'aide de polymorphismes de séquence pour quelques gènes candidats (gènes de Saccharose Synthéthase, de Caféine Synthéthase) ainsi qu'une approche ciblée sur le BAC111O18 qui colocalise avec un QTL d'intérêt de granulométrie et qui contient un gène homologue du gène *Ovate* de *Solanum Lycopersicum* impliqué dans la forme et la taille du fruit.

Conclusion

Ce chapitre nous a permis de compléter notre étude précédente par la comparaison de deux types de marqueurs possédant des modèles évolutifs et des taux de mutations fortement contrastés. Nous avons montré que ces deux types de marqueur peuvent être utilisés pour l'étude du déséquilibre de liaison à condition d'appliquer des corrections adaptées. Pour les applications en génétique d'association, il semble qu'il faudra moins de marqueurs microsatellites que de polymorphismes de séquences pour couvrir l'ensemble du génome. Néanmoins l'importance du nombre d'allèles des microsatellites est à prendre en compte et un nombre important de génotypes devra être utilisé avec ce type de marqueurs. En effet les SNPs étant des marqueurs bialléliques, le nombre de combinaisons alléliques possibles et de recombinaisons efficaces sont plus faibles que pour des marqueurs multialléliques tels que les

microsatellites. Nous pouvons donc observer un effet de dilution ne permettant pas de détecter l'ensemble des recombinaisons effectives à l'aide des polymorphismes de séquences. En revanche, ces marqueurs seront utiles pour des approches de types gènes candidats dans lesquelles on recherchera les polymorphismes responsables des variations de caractères d'intérêt (QTN ou Quantitative Trait Nucleotide).

Discussion et conclusion générale

La diversité génétique de *C. canephora*, une mine d'or pour la sélection

Nos études de diversité ont permis d'obtenir plusieurs résultats intéressants pour l'amélioration de *C. canephora* et d'autres espèces du genre.

La proximité génétique de *C. canephora* avec les autres espèces cultivées (*C. liberica* et *C. arabica*) permet d'espérer une généralisation des résultats obtenus sur notre espèce. Les régions génomiques et les gènes identifiés par cartographie génétique ou études d'association pourront être étudiés par synténie sur les autres espèces. De plus une introgression assistée par marqueurs de *C. canephora* sur *C. arabica* peut être facilement envisagée. Ce type d'approche pourrait permettre de suivre l'introgression de caractères tels que des résistances tout en conservant le patrimoine génétique global de l'espèce cible avec un gain de temps considérable pour des espèces pérennes, en imaginant des schémas d'introgression assistée par marqueurs.

La diversité génétique et phénotypique intraspécifique de *C. canephora* est très importante, fournissant une matière première de choix pour la sélection. Nous avons confirmé et précisé les résultats obtenus à l'aide d'autres marqueurs et permis de poser les bases d'une caractérisation rapide et efficace des collections ou de populations d'origine inconnue.

Une structuration en 7 groupes principaux a été mise en évidence à travers ce travail et des travaux annexes. Nous avons ainsi identifié les groupes Guinéens Pélézi et Autres Guinéens, et les groupes Congolais B, C, SG1, SG2 et Ougandais Sauvages (UW).

Les résultats du chapitre 3 permettent d'envisager la possibilité d'une origine unique des groupes B et SG2, et dans une moindre mesure UW, avec l'hypothèse que ces groupes correspondent en réalité à des prospections ponctuelles dans un centre de diversité primaire, correspondant grossièrement au bassin du Congo. Les groupes SG1 et C, quant à eux, bien que proches géographiquement, semblent plus distants génétiquement et pourraient correspondre à d'autres centres de diversité, de même que les Guinéens. Des prospections

complémentaires, notamment dans le bassin du Congo, seront nécessaires pour préciser cette hypothèse et valider ou infirmer la réalité des groupes de diversité identifiés.

Un autre résultat intéressant concerne le groupe SG2. En effet l'étude a porté sur des génotypes de la population INEAC de RCI composée de génotypes importés du Congo Belge par graines en 1935, ainsi que sur des génotypes cultivés d'Ouganda. Les génotypes cultivés d'Ouganda correspondent à une introduction de matériel du Congo Belge ayant transité par Java en 1910 (Thomas, 1935). Les indications de Cramer (1957) précisent que ce matériel avait été introduit à Java aux alentours de 1900, vraisemblablement avant qu'un travail de sélection important ait été mené puisque la culture de cette espèce n'avait tout au plus que quelques années. La proximité génétique observée entre ces origines différentes (35 ans entre les 2 exportations du Congo Belge) laisse penser que les génotypes observés sont proches génétiquement des génotypes sauvages originaires de RDC qui ont pu servir à la mise en place des premières plantations dans ce pays.

Nous confirmons donc que les populations cultivées que nous observons aujourd'hui sont proches génétiquement des populations originelles. Il n'y a pour l'instant aucune évidence de syndrome de domestication sur notre espèce.

La présence de 2 groupes au sein des Guinéens (Pélézi et autres Guinéens) différenciés génétiquement est un nouveau résultat. Cette spécificité génétique des Pélézi confirme leur originalité constatée sur le plan phénotypique et la valeur des hybrides ayant recours à cette population dans le schéma de SRR (Montagnon, 2000; Leroy, communication personnelle). D'autre part, nous avons montré l'existence d'une diversité non négligeable au sein de ce groupe, ainsi qu'une sous-structure complexe au niveau de la RCI.

Les ressources potentielles pour l'amélioration sont très importantes, avec des contrastes phénotypiques importants comme des tolérances à la sécheresse, la granulométrie, la teneur en caféine, les résistances ou tolérances aux maladies et insectes ayant pu être mis en évidence entre les différents groupes. Récemment une étude sur des caféiers ougandais a également mis en évidence la présence dans certaines origines de tolérance au Coffee Wilt Disease, maladie ré-émergente ravageant les cultures pouvant être une cible de choix rapide pour la sélection (Musoli, 2007).

Les stratégies d'amélioration intragroupe ont été fortement utilisées sur *C. canephora* par exemple en Indonésie ou au Congo Belge et ont permis un progrès génétique important (Montagnon *et al.*, 1998b). Néanmoins la valeur agronomique des hybrides intergroupes a été

mise en évidence en République de Côte d'Ivoire (Berthaud, 1986; Leroy, 1993; Montagnon, 2000) et de nombreux génotypes identifiés en plantation comme ayant une valeur particulièrement importante ont été caractérisés comme hybride intergroupes Congolais x Guinéens. Une première analyse datant de 2005 nous avait permis d'identifier l'un de ces génotypes (le clone 126) comme un hybride 3 voies ((SG1 x SG2) x G) (Cubry, 2005). La stratégie d'amélioration par Sélection Récurrente Réciproque menée en Côte d'Ivoire repose sur la création d'hybrides intergroupes et a déjà montré d'important succès et confirmé leur intérêt (Leroy *et al.*, 1993; Leroy *et al.*, 1997; Montagnon, 2000).

Choix de la stratégie pour la mise en place d'études d'association et relation avec le schéma de SRR

Le travail présenté ici a permis de caractériser quelques populations de *C. canephora* pour le déséquilibre de liaison, permettant de guider la mise en place d'études d'association. Nous avons utilisé ici un DL haplotypique, en complément d'études préliminaires réalisées avec un DL génotypique (Cubry, 2005). Les 2 approches donnent des résultats similaires bien que le DL haplotypique soit théoriquement plus puissant et donne des résultats plus précis. Cette similitude indique la robustesse de ces analyses, et peut se retrouver pour d'autres plantes comme le palmier à huile (Cochard, communication personnelle) ou la vigne (Barnaud *et al.*, 2006).

Quels modèles pour les études d'association sur Coffea canephora?

Nous avons vu au chapitre 4 qu'après correction des p-values des tests d'associations sur l'ensemble des génotypes et marqueurs par la méthode de Bonferroni, des valeurs importantes et significatives des 2 mesures de DL (D' et r^2) se retrouvaient aussi bien entre marqueurs non liés que liés, ne permettant pas de distinguer des associations reposant sur une liaison physique entre des marqueurs de celles créées par la structure. Afin de limiter les faux positifs dans les études d'association dus à cette confusion, divers modèles peuvent être utilisés pour s'affranchir de l'effet de la structure.

L'approche de type Genomic Control (Devlin *et al.*, 2001) que l'on peut décrire comme une adaptation du seuil de signification en fonction du nombre d'associations détectées entre des marqueurs non liés apparaît donc peu performante et entrainerait un nombre de faux négatifs beaucoup trop important. Cette critique est l'une des plus

importantes de ce modèle avancées par Yu *et al.* (2006). De plus cette approche estime que la structure a le même effet en tout point du génome (Yu & Buckler, 2006).

L'approche de Structured Association proposée par Pritchard *et al.* (2000b) semble plus efficace, néanmoins l'importance de l'apparentement dans nos populations comme nous le montre les arbres de diversité obtenus notamment pour la population Pélézi implique qu'une part de l'effet confondant de la structuration génétique n'est pas pris en compte dans ce modèle. Par conséquent, il semble que le modèle le mieux adapté à notre espèce et à nos populations soit le modèle mixte proposé par Yu *et al.* (2006). Cette approche a montré sa puissance et son meilleur contrôle des faux positifs par rapport aux autres méthodes sur données simulées.

Ces modèles d'études d'association sont de plus en plus performants et il semble que l'on arrive à un point critique de leur développement. Néanmoins une attention particulière doit être portée au choix des caractères étudiés et à leur répartition au sein de l'échantillon dans lequel sont menées les études d'association. En effet la correction de l'effet de la structure risque de gêner la détection d'associations pour des caractères qui auraient une répartition se superposant à la structure des populations (Camus-Kulandaivelu, 2006; Camus-Kulandaivelu *et al.*, 2006). En effet, certains caractères étant fortement corrélés à la structure en groupe de diversité pour notre espèce, comme par exemple la tolérance à la sécheresse (haute pour les Guinéens et les SG1, faible pour les autres groupes) ou la tolérance à la rouille orangée (haute pour les Congolais, faible pour les autres groupes), une réflexion sur les populations d'études utilisables pour ces caractères devra être menée.

Peut-on utiliser l'existant ? Importance de l'interaction GxE

L'une des principales idées sous-jacentes au développement des études d'association était de pouvoir s'adresser directement à des populations naturelles ou d'amélioration sans la nécessité de recourir à des croisements contrôlés et la mise en place de populations développées dans un but précis comme les populations de type F2, Backcross ou RILs. Cette possibilité revêt une importance toute particulière pour les espèces pérennes pour lesquelles l'obtention de telles populations demande un temps important. Une des questions qui se pose donc dans notre cas est la possibilité d'utiliser des populations d'amélioration ou des collections préexistantes. Nous avons particulièrement à disposition des populations et des collections présentes à Divo en Côte d'Ivoire, à Mukono en Ouganda et à Sinnamay en Guyane.

Chaque génotype des collections de travail de Divo en RCI est représenté par 4 boutures en ligne dont 2 plants écimés et 2 autres en croissance libre (Leroy, communication personnelle). Dans les études de génétique d'association et de génétique quantitative au sens large, la part de la variabilité phénotypique due à l'environnement doit être quantifiée. Il semble que le dispositif en place soit limitant pour la prise en compte de cette part de la variabilité due notamment au sol, à l'orientation par rapport au soleil, à la présence de pistes. Cette limitation est également valable pour l'estimation de l'interaction Génotype x Environnement. Cependant, s'agissant d'une culture pérenne, il existe un fort effet environnemental lié à l'année, à la pluviométrie, aux accidents de culture, qui ne sont pas liés à la place dans la parcelle et qui peut être pris en compte à travers une évaluation pluriannuelle des arbres.

Un premier test d'étude d'association a été réalisé sur des données issues de la collection de référence de Divo. Nous avons utilisé la moyenne de la granulométrie calculée sur 6 années pour des génotypes du groupe SG2 (21) et du groupe Autres Guinéens (57), pour un total de 78 génotypes. Les marqueurs utilisés sont ceux présentés au chapitre 4. La couverture du génome n'est pas optimale mais ce test permet de valider les potentialités des études d'association sur notre espèce. Nous avons utilisé le modèle mixte proposé par Yu *et al.* (2006), avec une matrice de structure calculée à l'aide du logiciel Structure 2.1 (Pritchard *et al.*, 2000a) et une matrice des apparentements calculée selon la méthode de Loiselle à l'aide du logiciel SpaGeDi (Hardy & vekemans, 2002). Sur les 5 associations significatives détectées, 4 se trouvent dans les intervalles de confiance de QTLs de granulométrie obtenus par cartographie génétique. Cet exemple nous permet donc, malgré les approximations notamment sur l'obtention des données phénotypiques, de valider ce type d'approche pour notre espèce.

Il est intéressant de constater l'obtention de résultats aussi concordants dans une collection de ce type, montrant sans équivoque l'intérêt et les possibilités de ces approches qui pourront être étendues à d'autres caractères s'il est possible d'effectuer des phénotypages sur plusieurs années.

La mise en place d'essais multi-sites de populations ou de core-collections pourront par la suite permettre, avec un dispositif statistique adapté, de travailler plus précisément en prenant en compte à la fois l'effet de l'environnement et l'interaction Génotype x Environnement, qui pour des espèces pérennes comme la nôtre présentent un enjeu particulièrement important.

Choix des populations

Nous avons montré le succès d'une approche d'étude d'association sur un échantillon composé majoritairement de génotypes Guinéens. Au vu de la structure du DL dans les 2 groupes considérés (SG2 et G), il semble probable que les associations détectées sont majoritairement dues aux Guinéens. Ce groupe de diversité, de même que les groupes C et SG1 pourrait être des cibles idéales pour des études d'association s'intéressant au génome entier. Ces groupes présentent en effet des étendues de DL moyennes à élevées pouvant permettre un crible génome entier avec un nombre raisonnable de marqueurs.

Le groupe de diversité SG2 pourra quant à lui être utilisé pour des approches gènes candidats. Des cores-collections moléculaires et morphologiques pourront être développées afin de limiter les coûts de maintien et de phénotypage tout en maximisant la diversité représentée. Cette approche, développée sur la vigne (Barnaud *et al.*, 2006; Le Cunff *et al.*, 2008) semble prometteuse et devrait permettre des avancées rapides dans la recherche d'associations marqueurs-caractères et le développement de la Sélection Assistée par Marqueurs.

Conclusion pour la sélection

Les études de diversité pourront amener une réflexion sur la conception des groupes utilisés dans le schéma de Sélection Récurrente et Réciproque, avec une restructuration possible des populations de travail. Une aide à la sélection par l'élimination précoce de génotypes redondants au cours des cycles pourra être mise en place comme première étape d'une stratégie de Sélection Assistée par Marqueurs. Des tests de croisements devront être réalisés entre les différents groupes décrits afin d'identifier les groupes les plus hétérotiques afin d'optimiser le schéma de sélection.

Il semble d'autre part qu'un premier crible d'études d'association sur des caractères à déterminisme génétique peu complexes et possédant une bonne héritabilité puisse être réalisé assez rapidement, sous la condition d'une densification de marquage sur certains groupes de liaison. De plus, des études de ce type pourraient être menées sur les essais mis en place dans le cadre du schéma de SRR où le dispositif statistique plus élaboré permettrait de mieux prendre en compte l'effet de l'environnement et de travailler sur des caractères au déterminisme plus complexe.

Références

Alvarez A E, van de Wiel C C M, Smulders M J M and Vosman B. **2001.** Use of microsatellites to evaluate genetic diversity and species relationships in the genus *Lycopersicon*. *Theor Appl Genet*, **114**:359-372.

Anthony F, Combes C, Astorga C, Bertrand B, Graziosi G and Lashermes P. **2002.** The origin of cultivated *Coffea arabica* L. varieties revealed by AFLP and SSR markers. *Theor Appl Genet*, **104**:894-900.

Barnaud A, Lacombe T and Doligez A. **2006.** Linkage disequilibrium in cultivated grapevine, *Vitis vinifera* L. *Theor Appl Genet*, **112**:708-716.

Berthaud J. **1986.** *Les ressources génétiques pour l'amélioration des caféiers africains diploïdes*, Paris. ORSTOM.

Berthaud J, Anthony F and Le Pierrès D. **1984.** Les caféiers de la Nana. Résultats des observations faites en collection en Côte d'Ivoire. *Café Cacao Thé*, **28**:3-13.

Berthaud J and Guillaumet J L. **1978.** Les caféiers sauvages en Centrafrique. Résultats d'une mission de prospection (janvier-février 1975). *Café Cacao Thé*, **22**:171-187.

Bouchet S. 2005. Diversité nucléotidique de cinq gènes impliqués dans la voie de biosynthèse du saccharose chez *Coffea canephora*: évolution moléculaire et implication pour la mise en place de la sélection assistée par marqueurs pour la qualité organoleptique. Master 2 recherche, Ecole Nationale Supérieure d'Agronomie de Rennes.

Camus-Kulandaivelu L. 2006. Évolution génomique du maïs lors de son adaptation aux conditions européennes. Thèse de Doctorat, Ecole Nationale Supérieure d'Agronomie de Montpellier.

Camus-Kulandaivelu L, Veyrieras J-B, Madur D, Combes V, Fourmann M, Barraud S, Dubreuill P, Gouesnard B, Manicacci D and Charcosset A. **2006.** Maize adaptation to temperate climate: relationship between population structure and polymorphism in the *Dwarf8* gene. *Genetics*, **172**:2449-2463.

Casasoli M. 2004. Cartographie comparée chez les Fagacées. Thèse de Doctorat, Bordeaux 1.

Chagné D. 2004. Développement de marqueurs moléculaires chez le pin maritime (Pinus pinaster Ait.) et cartographie génétique comparée des conifères. Thèse de Doctorat, Université Henri Poincaré.

Charrier A and Eskes A B. 2001. Coffee. In *Tropical plant breeding*, pp. 128-152.

Clauss M J, Cobban H and Mitchell-Olds T. **2002.** Cross-species microsatellite markers for elucidating population genetic structure in Arabidopsis and Arabis (Brassicaeae). *Mol Ecol*, **11**:591-601.

Combes M C, Andrzejewski S, Anthony F, Bertrand B, Rovelli P, Graziosi G and Lashermes P. **2000.** Characterization of microsatellite loci in Coffea arabica and related coffee species. *Mol Ecol*, **9**:1178-1180.

Coombs J A, Letcher B H and Nislow K H. 2007. CREATE 1.0 a Software to create and convert codominant molecular data. Available online at http://www.lsc.usgs.gov/CAFL/Ecology/Software.html.

Cramer P J S. **1957.** *A review of litterature of coffee research in Indonesia for about 1602 to 1945*, Turrialba, Costa-Rica. Interamerican Institute of Agricultural Science.

Cubry P. 2005. Analyse de la diversité et évaluation de déséquilibre de liaison chez quelques populations naturelles et cultivées de caféiers *Coffea canephora*. DEA, Université Montpellier II.

Cubry P, Musoli P, Legnaté H, Pot D, de Bellis F, Poncet V, Anthony F, Dufour M and Leroy T. **2008.** Diversity in coffee assessed with SSR markers: structure of the genus Coffea and perspectives for breeding. *Genome*, **51**:50-63.

Davis A, Govaerts R, Bridson D M and Stoffelen P. **2006.** An annotated taxonomic conspectus of the genus Coffea (Rubiaceae). *Botanical Journal of the Linnean Society*, **152**:465-512.

Devlin B and Roeder K. **1999.** Genomic control for association studies. *Biometrics*, **55**:997-1004.

Devlin B, Roeder K and Bacanu S A. **2001.** Unbiased methods for population-based association studies. *Genet Epidemiol*, **21**:273-284.

Dussert S, Lashermes P, Anthony F, Montagnon C, Trouslot P, Combes M C, Berthaud J, Noirot M and Hamon S. 1999. Le caféier, Coffea canephora. In *Diversité génétique des plantes tropicales cultivées*, pp. 175-194.

Eskes A B and Leroy T. 2004. Coffee selection and breeding. In *Coffee: growing, processing, sustainable production : A guidebook for growers, processors, traders, and researchers*, pp. 57-86.

Flint-Garcia S A, Thornsberry J M and Buckler E S t. **2003.** Structure of linkage disequilibrium in plants. *Annu Rev Plant Biol*, **54**:357-374.

Flint-Garcia S A, Thuillet A C, Yu J, Pressoir G, Romero S M, Mitchell S E, Doebley J, Kresovich S, Goodman M M and Buckler E S. **2005.** Maize association population: a high-resolution platform for quantitative trait locus dissection. *Plant J*, **44**:1054-1064.

Gao L Z, Zhang C H and Jia J Z. **2005.** Cross-species transferability of rice microsatellites in its wild relatives and the potential for conservation genetic studies. *Genetic Ressources and Crop Evolution*, **52**:931-940.

Gonzalez-Martinez S C, Huber D, Ersoz E, Davis J M and Neale D B. **2008.** Association genetics in Pinus taeda L. II. Carbon isotope discrimination. *Heredity*, **101**:19-26.

Gonzalez-Martinez S C, Wheeler N C, Ersoz E, Nelson C D and Neale D B. **2007.** Association genetics in Pinus taeda L. I. Wood property traits. *Genetics*, **175**:399-409.

Gouesnard B, Bataillon T M, Decoux G, Rozale C, Schoen D J and David J L. **2001.** MSTRAT: an algorithm for building germ plasm core collections by maximizing allelic or phenotypic richness. *J Hered*, **92**:93-94.

Hardy O J and vekemans X. **2002.** SPAGeDI: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular Ecology Notes*, **2**:618-620.

Hill W G and Robertson A. **1968.** Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, **38**:226-231.

INCO ICA4-CT-2001-10006 Final report, 2007. Development of a long term strategy based on genetic resistance ans agroecological approaches against Coffee Wilt Disease in Africa. Acronym: COWIDI

Ingvarsson P K. **2005.** Nucleotide polymorphism and linkage disequilibrium within and among natural populations of European aspen (Populus tremula L., Salicaceae). *Genetics*, **169**:945-953.

Lashermes P, Combes M C, Robert J, Trouslot P, D'Hont A, Anthony F and Charrier A. **1999.** Molecular characterisation and origin of the Coffea arabica L. genome. *Mol Gen Genet*, **261**:259-266.

Le Cunff L, Fournier-Level A, Laucou V, Vezzulli S, Lacombe T, Adam-Blondon A F, Boursiquot J M and This P. **2008.** Construction of nested genetic core collections to optimize the exploitation of natural diversity in Vitis vinifera L. subsp. sativa. *BMC Plant Biol*, **8**:31.

Leroy T, Montagnon C, Charrier A and Eskes A B. **1993.** Reciprocal Recurrent Selection applied to *Coffea canephora* Pierre. I. Characterization and evaluation of breeding populations and values of intergroup hybrids. *Euphytica*, **67**:113-125.

Leroy T, Montagnon C, Cilas C and Charrier A. **1994.** Reciprocal Recurrent Selection applied to *Coffea canephora* Pierre. II. Estimation of genetic parameters. *Euphytica*, **74**:121-128.

Leroy T, Montagnon C, Cilas C, Yapo A, Charmetant P and Eskes A B. **1997.** Reciprocal Recurrent Selection applied to *Coffea canephora* Pierre. III. Genetic gains and results of first intergroup crosses. *Euphytica*, **95**.

Lewontin R C. **1964.** The interactions of selection and linkage. I. General considerations; heterotic models. *Genetics*, **49**:49-67.

Lewontin R C and Kojima K I. **1960.** The evolutionary dynamics of complex polymorphisms. *Evolution*, **14**:458-472.

Liu J, Van Eck J, Cong B and Tanksley S D. **2002.** A new class of regulatory genes underlying the cause of pear-shaped tomato fruit. *Proc Natl Acad Sci U S A*, **99**:13302-13306.

Liu K and Muse S V. **2005.** PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics*, **21**:2128-2129.

Louarn J. 1992. La fertilité des hybrides interspécifiques et les relations génomiques entre caféiers diploïdes d'origine africaine (genre *Coffea* L., sous-genre *Coffea*). Thèse d'état, Paris XI.

Montagnon C. 2000. Optimisation des gains génétiques dans le schéma de sélection récurrente réciproque de *Coffea canephora* Pierre. Thèse de Doctorat, Ecole Nationale Supérieure Agronomique de Montpellier, France.

Montagnon C and Leroy T. **1993.** Réaction à la sécheresse de jeunes caféiers *Coffea canephora* de Côte-d'Ivoire appartenant à différents groupes génétiques. *Café Cacao Thé*, **37**:179-190.

Montagnon C, Leroy T and Eskes A B. **1998a.** Amélioration variétale de *Coffea canephora*. I. Critères et méthodes de sélection. *Plantation Recherche Développement*, **5**:18-28.

Montagnon C, Leroy T and Eskes A B. **1998b.** Amélioration variétale de *Coffea canephora*. II. Les programmes de sélection et leur résultats. *Plantation Recherche Développement*, **5**:89-98.

Montagnon C, Leroy T and Yapo A. **1992.** Diversité génotypique et phénotypique de quelques groupes de caféiers (*Coffea canephora* Pierre) en collection. *Café Cacao Thé*, **36**:187-198.

Musoli, P. 2007. Recherche de sources de résistance à la trachéomycose du caféier *Coffea canephora* Pierre, due à *Fusarium xylarioides* Steyaert en Ouganda. Thèse de Doctorat, Université de Montpellier 2, France.

Perrier X, Flori A and Bonnot F. 2003. Data analysis methods. In *Genetic diversity of cultivated tropical plants*, pp. 43-76.

Perrier X and Jacquemoud-Collet J P. 2006. DARwin software. http://darwin.cirad.fr/darwin.

Poncet V, Dufour M, Hamon P, Hamon S, de Kochko A and Leroy T. **2007.** Development of genomic microsatellite markers in Coffea canephora and their transferability to other coffee species. *Genome*, **50**:1156-1161.

Poncet V, Hamon P, Minier J, Carasco C, Hamon S and Noirot M. **2004.** SSR cross-amplification and variation within coffee trees (Coffea spp.). *Genome*, **47**:1071-1081.

Pritchard J K, Stephens M and Donnelly P. **2000a.** Inference of population structure using multilocus genotype data. *Genetics*, **155**:945-959.

Pritchard J K, Stephens M, Rosenberg N A and Donnelly P. **2000b.** Association mapping in structured populations. *Am J Hum Genet*, **67**:170-181.

Rafalski A and Morgante M. **2004.** Corn and humans: recombination and linkage disequilibrium in two genomes of similar size. *Trends Genet*, **20**:103-111.

Remington D L, Thornsberry J M, Matsuoka Y, Wilson L M, Whitt S R, Doebley J, Kresovich S, Goodman M M and Buckler E S t. **2001.** Structure of linkage

disequilibrium and phenotypic associations in the maize genome. *Proc Natl Acad Sci U S A*, **98**:11479-11484.

Stephens M and Donnelly P. **2003.** A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet*, **73**:1162-1169.

Stephens M and Scheet P. **2005.** Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am J Hum Genet*, **76**:449-462.

Stephens M, Smith N J and Donnelly P. **2001.** A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet*, **68**:978-989.

Thomas A S. **1935.** Types of Robusta coffee and their selection in Uganda. *The East African Agricultural Journal*, **1**:193-198.

Thornsberry J M, Goodman M M, Doebley J, Kresovich S, Nielsen D and Buckler E S t. **2001.** Dwarf8 polymorphisms associate with variation in flowering time. *Nat Genet*, **28**:286-289.

Veyrieras J-B. 2006. Etude du déterminisme génétique de caractères quantitatifs chez les végétaux. Thèse de Doctorat, Institut National Agronomique Paris-Grignon.

Wang W Y S, Baratt B J, Clayton D G and Todd J A. **2005.** Genome-wide association studies: Theoretical and practical concerns. *Nature Review Genetics*, **6**:109-118.

Weir B S. **1996.** *Genetic Data Analysis II: Methods for discrete population genetic data*. Sinauer Associates.

Yu J and Buckler E S. **2006.** Genetic association mapping and genome organization of maize. *Curr Opin Biotechnol*, **17**:155-160.

Yu J, Pressoir G, Briggs W H, Vroh Bi I, Yamasaki M, Doebley J F, McMullen M D, Gaut B S, Nielsen D M, Holland J B, Kresovich S and Buckler E S. **2006.** A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet*, **38**:203-208.

Yu K, Park S J and Poysa V. **1999.** Abundance and variation of microsatellite DNA sequences in beans (*Phaseolus* and *Vigna*). *Genome*, **42**:27-34.

Zhu C, Gore M, Buckler E and Yu J. **2008.** Status and prospects of association mapping in plants. *The Plant Genome*, **1**:5-20.

Annexes

Chapitre 2

Tableaux supplémentaires pour « Diversity in coffee assessed with SSR markers: structure of the genus Coffea and perspectives for breeding »

A.2.1 : Tableau supplémentaire 1 (pourcentage d'amplification par marqueurs pour les 15 espèces de l'étude)

A.2.2 : Tableau supplémentaire 2 (liste des allèles spécifiques dans les 15 espèces)

A.2.3 : Tableau supplémentaire 3 (répartition des allèles spécifiques par espèce)

A.2.4 : Tableau supplémentaire 4 (statistiques descriptives pour les 60 marqueurs calculées sur l'échantillon global et chacune des 15 espèces)

Table S1. Percentage of amplification (availability) per marker for the 15 species of the study

Marker	overall	arabica	canephora	liberica	congensis	anthonyi	bertrandii	brevipes	eugenioides	humilis	milloti	pseudozanguebariae	racemosa	salvatrix	sessiliflora	stenophylla
DL003	0.810	0.750	0.778	0.857	0.600	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.000	1.000	1.000
DL010	0.762	1.000	1.000	0.714	1.000	1.000	1.000	1.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	1.000
DL011	0.833	1.000	1.000	1.000	1.000	0.000	0.000	1.000	1.000	1.000	0.000	0.000	0.000	0.000	0.667	1.000
DL013	0.714	0.875	0.556	0.714	0.600	1.000	1.000	1.000	1.000	1.000	0.000	1.000	1.000	1.000	0.333	1.000
DL020	0.976	1.000	0.889	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
DL025	0.952	1.000	1.000	0.857	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.667	1.000
DL026	0.857	1.000	0.889	1.000	1.000	1.000	0.000	1.000	1.000	0.000	1.000	0.000	0.000	1.000	1.000	0.000
DL032	0.857	0.750	1.000	0.714	0.800	1.000	1.000	1.000	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000
SSR16	0.786	0.750	0.778	0.429	0.800	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
SSR14	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
SSR15	0.833	0.875	0.889	0.429	1.000	1.000	1.000	1.000	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000
SSR17	0.786	1.000	0.222	1.000	1.000	1.000	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	0.667	1.000
SSR1	0.548	1.000	1.000	0.000	1.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
SSR3	0.929	1.000	0.889	0.857	0.800	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
SSR4	0.976	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.667	1.000
257	0.810	0.625	0.889	0.571	0.800	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
305	0.810	0.875	0.556	0.714	1.000	1.000	1.000	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
327	0.833	0.750	0.778	0.714	0.800	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
329	0.905	1.000	0.889	1.000	0.800	1.000	1.000	1.000	1.000	0.000	0.000	1.000	1.000	1.000	1.000	1.000
334	0.952	1.000	1.000	0.857	0.800	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
341	0.429	0.875	0.444	0.571	0.400	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
350	0.929	1.000	0.778	0.857	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
351	0.857	1.000	0.778	0.714	0.800	1.000	1.000	1.000	1.000	1.000	0.000	1.000	1.000	1.000	1.000	1.000
355	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
356	0.738	0.750	1.000	0.143	0.600	1.000	1.000	1.000	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000
358	0.643	0.750	1.000	0.714	0.200	1.000	1.000	1.000	0.000	1.000	1.000	0.000	0.000	0.000	0.000	1.000
SSR9	0.714	0.875	1.000	0.571	0.800	0.000	1.000	1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.667	1.000
SSR10	0.310	0.625	0.889	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
360	0.452	0.000	1.000	0.714	0.600	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	1.000
364	0.881	1.000	1.000	0.714	0.800	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.333	1.000
367	0.905	1.000	1.000	0.857	0.800	0.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
368	0.881	0.875	1.000	0.571	0.800	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
371	0.905	1.000	0.889	1.000	1.000	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.667	0.000
384	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
388	0.810	1.000	1.000	0.286	1.000	0.000	1.000	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	0.000
392	0.905	1.000	0.889	1.000	1.000	1.000	0.000	0.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

394	0.952	1.000	1.000	0.857	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.000	1.000	1.000	1.000	1.000
395	0.714	0.875	0.778	0.714	1.000	0.000	1.000	1.000	0.000	0.000	1.000	1.000	1.000	1.000	0.000	0.000
429	0.857	1.000	1.000	0.714	0.800	0.000	1.000	1.000	0.000	1.000	1.000	1.000	1.000	1.000	0.667	1.000
442	0.381	0.750	0.778	0.000	0.600	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
445	0.833	0.625	1.000	0.857	0.800	1.000	0.000	1.000	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000
456	0.405	0.875	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.333	0.000
460	0.857	1.000	0.889	0.714	1.000	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.667	1.000
461	0.905	1.000	1.000	1.000	1.000	1.000	0.000	1.000	1.000	1.000	0.000	1.000	1.000	0.000	0.667	1.000
463	0.905	0.875	1.000	0.714	0.800	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
471	0.738	0.875	0.778	0.857	0.800	0.000	0.000	0.000	0.000	1.000	0.000	1.000	1.000	1.000	0.667	1.000
472	0.714	0.875	0.889	0.429	0.800	1.000	1.000	1.000	1.000	0.000	1.000	0.000	1.000	0.000	0.667	0.000
477	0.905	1.000	0.667	1.000	1.000	1.000	1.000	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
495	0.976	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.000	1.000	1.000	1.000	1.000
SSR5	0.905	1.000	1.000	0.714	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.667	1.000
501	0.881	1.000	1.000	1.000	1.000	1.000	0.000	1.000	1.000	1.000	1.000	0.000	0.000	0.000	0.667	1.000
753	0.762	1.000	0.889	0.429	0.800	0.000	1.000	1.000	1.000	1.000	1.000	1.000	0.000	1.000	0.667	0.000
755	0.857	0.750	1.000	0.714	0.800	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.667	1.000
774	0.952	1.000	1.000	1.000	1.000	1.000	0.000	1.000	1.000	1.000	0.000	1.000	1.000	1.000	1.000	1.000
779	0.929	0.750	0.889	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
782	0.833	0.500	0.889	0.714	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
790	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
809	0.786	0.750	1.000	0.714	0.800	1.000	0.000	1.000	1.000	1.000	0.000	1.000	1.000	1.000	0.333	1.000
837	0.786	0.875	0.778	0.714	0.800	1.000	1.000	1.000	0.000	1.000	0.000	1.000	1.000	1.000	0.667	1.000
838	0.810	1.000	1.000	0.714	0.400	1.000	1.000	1.000	1.000	1.000	1.000	0.000	1.000	1.000	0.333	1.000
Mean	0.815	0.890	0.898	0.729	0.830	0.767	0.700	0.883	0.733	0.750	0.717	0.733	0.800	0.733	0.722	0.817
Mean†	0.816	0.889	0.898	0.728	0.829	0.766	0.701	0.884	0.733	0.750	0.716	0.734	0.800	0.734	0.722	0.817
SD†	0.020	0.023	0.020	0.036	0.030	0.055	0.060	0.042	0.057	0.056	0.060	0.058	0.052	0.058	0.046	0.050
2.5% l.b.†	0.774	0.842	0.857	0.655	0.767	0.650	0.583	0.800	0.617	0.633	0.600	0.617	0.700	0.617	0.628	0.717
97.5% l.b.†	0.852	0.931	0.935	0.795	0.887	0.867	0.817	0.950	0.833	0.850	0.833	0.833	0.900	0.833	0.811	0.900

† Mean values are calculated over 5000 bootstrap iterations and based only on markers with no missing data for the considered species

SD is Standard Deviation, 2.5% l.b. and 97.5% u.b. are lower and upper boundaries of the 95% confidence interval.

Table S2. List of private alleles in the 15 species

Locus	Allele size	Private to
DL003	173	milloti
DL010	158	canephora
DL010	170	humilis
DL010	172	moloundou
DL010	173	bertrandii
DL010	179	canephora
DL010	184	arabica
DL011	273	liberica
DL011	288	canephora
DL011	291	brevipes
DL013	274	stenophylla
DL013	282	bertrandii
DL013	286	arabica
DL013	288	canephora
DL013	300	liberica
DL013	302	racemosa
DL020	235	sessiliflora
DL020	253	canephora
DL020	255	canephora
DL020	257	canephora
DL025	212	liberica
DL026	131	sessiliflora
DL026	132	sessiliflora
DL026	135	liberica
DL026	140	canephora
DL026	141	canephora
DL026	144	salvatrix
DL026	145	milloti
DL032	249	pseudozanguebariae
DL032	261	liberica
SSR14	202	moloundou
SSR14	203	canephora
SSR14	212	pseudozanguebariae
SSR14	217	racemosa
SSR15	141	stenophylla
SSR15	162	sessiliflora
SSR17	141	liberica
SSR17	157	congensis
SSR17	161	liberica
SSR17	169	bertrandii
SSR3	141	pseudozanguebariae
SSR4	128	stenophylla
257	119	humilis
257	131	stenophylla
257	135	liberica
257	143	liberica
257	151	arabica
257	153	arabica
257	157	sessiliflora
257	159	milloti
257	169	bertrandii
305	151	bertrandii
305	157	moloundou
305	161	pseudozanguebariae
305	163	canephora
327	200	congensis
327	201	canephora
327	202	canephora
327	209	canephora
327	210	canephora
327	214	brevipes
327	217	arabica
327	219	arabica
329	242	congensis
329	246	liberica
329	250	salvatrix
329	256	congensis
329	258	canephora
329	260	canephora
329	274	stenophylla
334	104	canephora
341	130	canephora
341	132	liberica
341	150	brevipes
350	313	congensis
350	315	congensis
350	329	liberica
350	331	arabica
350	333	arabica
351	315	eugenioides
351	327	moloundou
351	341	humilis
355	160	pseudozanguebariae
355	166	salvatrix
355	196	congensis
355	202	liberica
355	210	canephora
355	212	moloundou
356	175	brevipes
356	185	sessiliflora
356	189	sessiliflora
356	199	canephora
356	203	milloti
356	209	moloundou
358	270	canephora
358	284	canephora
358	286	stenophylla
358	290	milloti
358	304	canephora
SSR9	89	canephora
SSR9	95	bertrandii
SSR9	97	liberica
SSR9	101	liberica
SSR9	105	eugenioides
SSR9	113	liberica
SSR10	117	canephora
SSR10	121	canephora
SSR10	123	canephora
360	199	liberica
360	207	liberica
360	209	liberica
360	219	liberica
360	223	canephora
360	225	canephora
360	227	canephora
360	233	congensis
364	96	canephora
364	102	canephora
364	108	canephora
367	179	liberica
367	199	brevipes
367	201	arabica
367	213	canephora
368	162	humilis
368	164	brevipes
368	172	canephora
368	175	canephora
368	178	canephora
368	182	congensis
368	184	canephora
368	186	liberica
368	194	bertrandii

368	218	sessiliflora
371	314	canephora
371	316	liberica
371	320	brevipes
371	328	congensis
371	336	congensis
384	279	congensis
388	339	canephora
388	343	pseudozanguebariae
388	351	canephora
388	359	racemosa
388	399	bertrandii
388	403	sessiliflora
388	405	sessiliflora
388	409	milloti
388	421	humilis
388	423	liberica
388	433	liberica
392	245	liberica
392	251	sessiliflora
392	252	sessiliflora
392	257	liberica
392	269	liberica
392	277	canephora
394	134	milloti
394	138	sessiliflora
394	140	sessiliflora
394	156	liberica
394	158	stemophylla
394	162	humilis
394	164	liberica
395	106	liberica
395	120	congensis
395	128	canephora
395	132	canephora
395	136	canephora
395	144	arabica
395	150	arabica
395	152	arabica
395	156	milloti
429	158	racemosa
429	164	brevipes
429	180	canephora
429	183	humilis
429	186	brevipes
429	189	liberica
429	191	congensis
429	194	canephora
429	198	liberica
429	200	canephora
429	202	sessiliflora
429	204	canephora
442	220	canephora
442	226	canephora
442	232	canephora
442	234	congensis
442	236	canephora
442	248	canephora
445	282	sessiliflora
445	288	canephora
445	300	milloti
445	308	racemosa
456	263	canephora
456	285	canephora
456	291	canephora
456	293	canephora
456	295	canephora
456	297	canephora
456	299	canephora
456	303	canephora
456	309	canephora
460	344	liberica
460	354	brevipes
460	360	canephora
460	366	canephora
460	370	congensis
460	376	congensis
460	378	congensis
460	380	congensis
460	382	liberica
460	392	stemophylla
460	394	liberica
460	396	arabica
460	408	liberica
461	88	liberica
461	92	liberica
461	106	liberica
461	112	arabica
461	118	canephora
463	250	pseudozanguebriae
463	252	sessiliflora
471	303	congensis
471	319	stemophylla
471	329	liberica
472	318	bertrandii
472	320	milloti
472	322	sessiliflora
472	324	sessiliflora
472	334	liberica
472	344	canephora
472	348	arabica
472	350	canephora
477	272	liberica
477	276	stemophylla
477	278	canephora
477	282	canephora
477	288	canephora
477	294	milloti
477	300	salvatrix
495	222	racemosa
495	228	canephora
495	230	canephora
SSR5	124	congensis
SSR5	140	pseudozanguebariae
SSR5	145	liberica
SSR5	151	canephora
SSR5	153	canephora
SSR5	174	congensis
501	141	congensis
501	143	moloundou
501	151	canephora
501	155	canephora
501	157	liberica
501	161	arabica
501	165	canephora
501	167	liberica
501	169	canephora
501	171	liberica
501	175	canephora
753	285	canephora
753	291	milloti
753	295	salvatrix
753	297	milloti
753	299	pseudozanguebariae
753	303	liberica
753	307	sessiliflora
753	319	liberica
755	171	canephora
755	181	canephora
755	186	stemophylla
755	191	racemosa
755	195	canephora

774	216	humilis
774	228	canephora
774	232	brevipes
774	238	brevipes
774	244	arabica
782	136	sessiliflora
782	140	congensis
782	150	brevipes
790	126	canephora
790	136	canephora
790	142	stemophylla
790	148	racemosa
790	150	canephora
790	154	canephora
790	160	arabica
809	138	liberica
809	144	liberica
809	146	canephora
809	156	liberica
837	107	congensis
837	117	arabica
837	121	congensis
837	131	canephora
838	117	milloti
838	123	canephora
838	127	canephora
838	133	congensis
838	137	arabica
838	139	brevipes
838	143	brevipes

Table S3. Repartition of private alleles per species

Species	Number of private alleles	Percentage of private alleles per species upon the number of private alleles	Percentage of private alleles per species upon the total number of alleles
canephora	95	31.25	14.66
brevipes	15	4.93	2.31
congensis	27	8.88	4.17
arabica	20	6.58	3.09
eugenioides	2	0.66	0.31
anthonyi	0	0.00	0.00
liberica	52	17.11	8.02
milloti	13	4.28	2.01
bertrandii	8	2.63	1.23
stenophylla	12	3.95	1.85
sessiliflora	21	6.91	3.24
humilis	7	2.30	1.08
racemosa	8	2.63	1.23
pseudozanguebariae	9	2.96	1.39
salvatrix	5	1.64	0.77
Mean	19.6	6.45	3.02

Table S4. Summary statistics calculated for the 60 SSR markers over the global sample and for each of the 15 species. Allele number (Allele No), Gene Diversity (GeneDiv.) and Observed Heterozigosity (ObsHet.) are presented.

Disponible online sur le site de Genome

Chapitre 3

Tableaux supplémentaires pour « Diversity and population structure of Coffea canephora (Rubiaceae) assessed by microsatellites. »

A.3.1 : Tableau supplémentaire 1 (liste des génotypes utilisés dans cette étude)

A.3.2 : Tableau supplémentaire 2 (F_{st} deux à deux pour les différents niveaux d'étude de la structure)

A.3.3 : Tableau supplémentaire 3 (AMOVAs basée sur les F_{st} et F-statistiques dérivées pour les différents niveaux d'étude de la structure)

A.3.4 : Tableau supplémentaire 4 (F_{is} par population pour les différents niveaux d'étude de la structure)

Table S1: Complete list of genotypes used in this study

Work Code	Population	Putative group based on collection data	Original collection name	Collection or Provider	Putative origin	Type
sp43	Nemaya		T3561 (Nemaya)	CATIE	DRC	unknown
sp44	Nemaya		T3761 (Nemaya)	CATIE	Central africa via Indonesia	unknown
br28	Brazil		apoatá E5	Brasil	unknown	unknown
br29	Brazil		apoatá H9	Brasil	unknown	unknown
br30	Brazil		L161	Brasil	unknown	unknown
br31	Brazil		G30	Brasil	unknown	unknown
br32	Brazil		K82	Brasil	unknown	unknown
c1001	Libengue	B	02519	CNRA - Divo	CAR near Libengé	wild
c1002	Libengue	B	02525	CNRA - Divo	CAR near Libengé	wild
c1003	Libengue	B	02536	CNRA - Divo	CAR near Libengé	wild
c1004	Libengue	B	02544	CNRA - Divo	CAR near Libengé	wild
c1005	Libengue	B	02556	CNRA - Divo	CAR near Libengé	wild
c1006	Libengue	B	02569	CNRA - Divo	CAR near Libengé	wild
c1007	Libengue	B	02572	CNRA - Divo	CAR near Libengé	wild
c1008	Libengue	B	02576	CNRA - Divo	CAR near Libengé	wild
c1009	Libengue	B	02589	CNRA - Divo	CAR near Libengé	wild
c1010	Libengue	B	02596	CNRA - Divo	CAR near Libengé	wild
c1011	Libengue	B	02518	CNRA - Divo	CAR near Libengé	wild
c1012	Libengue	B	02520	CNRA - Divo	CAR near Libengé	wild
c1013	Libengue	B	02521	CNRA - Divo	CAR near Libengé	wild
c1014	Libengue	B	02522	CNRA - Divo	CAR near Libengé	wild
c1015	Libengue	B	02524	CNRA - Divo	CAR near Libengé	wild
c1016	Libengue	B	02526	CNRA - Divo	CAR near Libengé	wild
c1017	Libengue	B	02527	CNRA - Divo	CAR near Libengé	wild
c1018	Libengue	B	02528	CNRA - Divo	CAR near Libengé	wild
c1019	Libengue	B	02529	CNRA - Divo	CAR near Libengé	wild
c1020	Libengue	B	02532	CNRA - Divo	CAR near Libengé	wild
c1021	Libengue	B	02533	CNRA - Divo	CAR near Libengé	wild
c1022	Libengue	B	02534	CNRA - Divo	CAR near Libengé	wild
c1023	Libengue	B	02538	CNRA - Divo	CAR near Libengé	wild

c1024	Libengue	B	02543	CNRA - Divo	CAR near Libengé	wild
c1025	Libengue	B	02550	CNRA - Divo	CAR near Libengé	wild
c1026	Libengue	B	02551	CNRA - Divo	CAR near Libengé	wild
c1028	Libengue	B	02561	CNRA - Divo	CAR near Libengé	wild
c1030	Libengue	B	02571	CNRA - Divo	CAR near Libengé	wild
c1033	Libengue	B	02579	CNRA - Divo	CAR near Libengé	wild
c1034	Libengue	B	02580	CNRA - Divo	CAR near Libengé	wild
c1035	Libengue	B	02583	CNRA - Divo	CAR near Libengé	wild
c1037	Libengue	B	02590	CNRA - Divo	CAR near Libengé	wild
c1038	Libengue	B	02592	CNRA - Divo	CAR near Libengé	wild
c1039	Libengue	B	02593	CNRA - Divo	CAR near Libengé	wild
c2001	INEAC2	SG2	027	CNRA - Divo	INEAC (Yangambi)	improved
c2002	INEAC2	SG2	032	CNRA - Divo	INEAC (Yangambi)	improved
c2003	INEAC2	SG2	036	CNRA - Divo	INEAC (Yangambi)	improved
c2004	INEAC2	SG2	042	CNRA - Divo	INEAC (Yangambi)	improved
c2006	INEAC2	SG2	049	CNRA - Divo	INEAC (Yangambi)	improved
c2007	INEAC2	SG2	066	CNRA - Divo	INEAC (Yangambi)	improved
c2008	INEAC2	SG2	067	CNRA - Divo	INEAC (Yangambi)	improved
c2013	INEAC2	SG2	A23	CNRA - Divo	INEAC (Yangambi)	improved
c2014	INEAC2	SG2	A24	CNRA - Divo	INEAC (Yangambi)	improved
c2015	INEAC2	SG2	092	CNRA - Divo	INEAC (Yangambi)	improved
c2016	INEAC2	SG2	02	CNRA - Divo	INEAC (Yangambi)	improved
c2017	INEAC2	SG2	037	CNRA - Divo	INEAC (Yangambi)	improved
c2018	INEAC2	SG2	047	CNRA - Divo	INEAC (Yangambi)	improved
c3001	Niaouli	SG1	NIAOULI1	CNRA - Divo	Gabon or Congo via Benin	cultivated
c3002	Niaouli	SG1	NIAOULI13	CNRA - Divo	Gabon or Congo via Benin	cultivated
c3003	Niaouli	SG1	NIAOULI14	CNRA - Divo	Gabon or Congo via Benin	cultivated
c3004	Niaouli	SG1	NIAOULI17	CNRA - Divo	Gabon or Congo via Benin	cultivated
c3005	Niaouli	SG1	NIAOULI20	CNRA - Divo	Gabon or Congo via Benin	cultivated
c3007	Niaouli	SG1	NIAOULI3	CNRA - Divo	Gabon or Congo via Benin	cultivated
c3008	Niaouli	SG1	NIAOULI6	CNRA - Divo	Gabon or Congo via Benin	cultivated
c3009	Niaouli	SG1	NIAOULI9	CNRA - Divo	Gabon or Congo via Benin	cultivated
c4001	Nana	C	Na003	CNRA - Divo	CAR	wild
c4002	Nana	C	Na007	CNRA - Divo	CAR	wild
c4003	Nana	C	Na025	CNRA - Divo	CAR	wild

c4004	Nana	C	Na033	CNRA - Divo	CAR	wild
c4005	Nana	C	Na035	CNRA - Divo	CAR	wild
c4006	Nana	C	Na036	CNRA - Divo	CAR	wild
c4007	Nana	C	Na040	CNRA - Divo	CAR	wild
c4008	Nana	C	Na054	CNRA - Divo	CAR	wild
c4009	Nana	C	Na090	CNRA - Divo	CAR	wild
c4010	Nana	C	Na097	CNRA - Divo	CAR	wild
c4011	Nana	C	Na001	CNRA - Divo	CAR	wild
c4012	Nana	C	Na002	CNRA - Divo	CAR	wild
c4013	Nana	C	Na004	CNRA - Divo	CAR	wild
c4014	Nana	C	Na005	CNRA - Divo	CAR	wild
c4015	Nana	C	Na006	CNRA - Divo	CAR	wild
c4016	Nana	C	Na008	CNRA - Divo	CAR	wild
c4017	Nana	C	Na009	CNRA - Divo	CAR	wild
c4018	Nana	C	Na010	CNRA - Divo	CAR	wild
c4019	Nana	C	Na011	CNRA - Divo	CAR	wild
c4020	Nana	C	Na012	CNRA - Divo	CAR	wild
c4021	Nana	C	Na013	CNRA - Divo	CAR	wild
c4024	Nana	C	Na018	CNRA - Divo	CAR	wild
c4026	Nana	C	Na020	CNRA - Divo	CAR	wild
c4027	Nana	C	Na021	CNRA - Divo	CAR	wild
c4028	Nana	C	Na022	CNRA - Divo	CAR	wild
c4029	Nana	C	Na023	CNRA - Divo	CAR	wild
c4030	Nana	C	Na026	CNRA - Divo	CAR	wild
c4031	Nana	C	Na027	CNRA - Divo	CAR	wild
c4032	Nana	C	Na028	CNRA - Divo	CAR	wild
c4033	Nana	C	Na029	CNRA - Divo	CAR	wild
c4034	Nana	C	Na030	CNRA - Divo	CAR	wild
c4035	Nana	C	Na031	CNRA - Divo	CAR	wild
c4036	Nana	C	Na032	CNRA - Divo	CAR	wild
c4038	Nana	C	Na041	CNRA - Divo	CAR	wild
c4039	Nana	C	Na043	CNRA - Divo	CAR	wild
c4040	Nana	C	Na044	CNRA - Divo	CAR	wild
c5002	INEAC7	SG2	006	CNRA - Divo	INEAC (Yangambi)	improved
c5003	INEAC7	SG2	008	CNRA - Divo	INEAC (Yangambi)	improved

c5004	INEAC7	SG2	015	CNRA - Divo	INEAC (Yangambi)	improved
c5005	INEAC7	SG2	022	CNRA - Divo	INEAC (Yangambi)	improved
c5006	INEAC7	SG2	024	CNRA - Divo	INEAC (Yangambi)	improved
g1001	Pelezi	G	02362	CNRA - Divo	Ivory Coast	wild
g1002	Pelezi	G	02363	CNRA - Divo	Ivory Coast	wild
g1003	Pelezi	G	02364	CNRA - Divo	Ivory Coast	wild
g1005	Pelezi	G	02367	CNRA - Divo	Ivory Coast	wild
g1006	Pelezi	G	02369	CNRA - Divo	Ivory Coast	wild
g1007	Pelezi	G	02370	CNRA - Divo	Ivory Coast	wild
g1008	Pelezi	G	02371	CNRA - Divo	Ivory Coast	wild
g1009	Pelezi	G	02374	CNRA - Divo	Ivory Coast	wild
g1010	Pelezi	G	02375	CNRA - Divo	Ivory Coast	wild
g1011	Pelezi	G	02376	CNRA - Divo	Ivory Coast	wild
g1013	Pelezi	G	02378	CNRA - Divo	Ivory Coast	wild
g1014	Pelezi	G	02379	CNRA - Divo	Ivory Coast	wild
g1015	Pelezi	G	02380	CNRA - Divo	Ivory Coast	wild
g1016	Pelezi	G	02381	CNRA - Divo	Ivory Coast	wild
g1017	Pelezi	G	02382	CNRA - Divo	Ivory Coast	wild
g1018	Pelezi	G	02384	CNRA - Divo	Ivory Coast	wild
g1019	Pelezi	G	02385	CNRA - Divo	Ivory Coast	wild
g1020	Pelezi	G	02386	CNRA - Divo	Ivory Coast	wild
g1021	Pelezi	G	02387	CNRA - Divo	Ivory Coast	wild
g1022	Pelezi	G	02388	CNRA - Divo	Ivory Coast	wild
g1023	Pelezi	G	02392	CNRA - Divo	Ivory Coast	wild
g1024	Pelezi	G	02393	CNRA - Divo	Ivory Coast	wild
g1025	Pelezi	G	02395	CNRA - Divo	Ivory Coast	wild
g1026	Pelezi	G	02399	CNRA - Divo	Ivory Coast	wild
g1027	Pelezi	G	02401	CNRA - Divo	Ivory Coast	wild
g1028	Pelezi	G	02360	CNRA - Divo	Ivory Coast	wild
g1029	Pelezi	G	02361	CNRA - Divo	Ivory Coast	wild
g1030	Pelezi	G	02366	CNRA - Divo	Ivory Coast	wild
g1031	Pelezi	G	02372	CNRA - Divo	Ivory Coast	wild
g1032	Pelezi	G	02377	CNRA - Divo	Ivory Coast	wild
g1033	Pelezi	G	02383	CNRA - Divo	Ivory Coast	wild
g1034	Pelezi	G	02389	CNRA - Divo	Ivory Coast	wild

g1035	Pelezi	G	02390	CNRA - Divo	Ivory Coast	wild
g1036	Pelezi	G	02391	CNRA - Divo	Ivory Coast	wild
g1037	Pelezi	G	02396	CNRA - Divo	Ivory Coast	wild
g1038	Pelezi	G	02398	CNRA - Divo	Ivory Coast	wild
g2001	Pine	G	02801	CNRA - Divo	Guinea	wild
g2002	Pine	G	02802	CNRA - Divo	Guinea	wild
g2003	Pine	G	02803	CNRA - Divo	Guinea	wild
g2004	Pine	G	02804	CNRA - Divo	Guinea	wild
g2005	Pine	G	02805	CNRA - Divo	Guinea	wild
g2006	Pine	G	02806	CNRA - Divo	Guinea	wild
g2008	Pine	G	02809	CNRA - Divo	Guinea	wild
g2009	Pine	G	02810	CNRA - Divo	Guinea	wild
g2010	Pine	G	02812	CNRA - Divo	Guinea	wild
g2011	Pine	G	02814	CNRA - Divo	Guinea	wild
g2012	Pine	G	02815	CNRA - Divo	Guinea	wild
g2013	Pine	G	02817	CNRA - Divo	Guinea	wild
g2014	Pine	G	02818	CNRA - Divo	Guinea	wild
g2015	Pine	G	02819	CNRA - Divo	Guinea	wild
g2016	Pine	G	02821	CNRA - Divo	Guinea	wild
g2017	Pine	G	02822	CNRA - Divo	Guinea	wild
g2018	Pine	G	02823	CNRA - Divo	Guinea	wild
g2020	Pine	G	02826	CNRA - Divo	Guinea	wild
g2021	Pine	G	02827	CNRA - Divo	Guinea	wild
g2023	Pine	G	02830	CNRA - Divo	Guinea	wild
g2024	Pine	G	02831	CNRA - Divo	Guinea	wild
g2025	Pine	G	02833	CNRA - Divo	Guinea	wild
g2026	Pine	G	02834	CNRA - Divo	Guinea	wild
g2027	Pine	G	02835	CNRA - Divo	Guinea	wild
g2029	Pine	G	02839	CNRA - Divo	Guinea	wild
g2030	Pine	G	02840	CNRA - Divo	Guinea	wild
g2031	Pine	G	02841	CNRA - Divo	Guinea	wild
g2032	Pine	G	02842	CNRA - Divo	Guinea	wild
g2033	Pine	G	02843	CNRA - Divo	Guinea	wild
g2034	Pine	G	02844	CNRA - Divo	Guinea	wild
g2035	Pine	G	02845	CNRA - Divo	Guinea	wild

g3001	Cultivated Ivorians	G	155	CNRA - Divo	Ivory Coast	cultivated/improved
g3003	Cultivated Ivorian	G	AYA01	CNRA - Divo	Ivory Coast	cultivated/improved
g3004	Cultivated Ivorian	G	DJE6	CNRA - Divo	Ivory Coast	cultivated/improved
g3005	Cultivated Ivorian	G	KAL81	CNRA - Divo	Ivory Coast	cultivated/improved
g3007	Cultivated Ivorian	G	KB9	CNRA - Divo	Ivory Coast	cultivated/improved
g3008	Cultivated Ivorian	G	ZAN60	CNRA - Divo	Ivory Coast	cultivated/improved
g3009	Cultivated Ivorian	G	150	CNRA - Divo	Ivory Coast	cultivated/improved
g3010	Cultivated Ivorian	Hybrid	243	CNRA - Divo	Ivory Coast	cultivated/improved
g3011	Cultivated Ivorian	Hybrid	418	CNRA - Divo	Ivory Coast	cultivated/improved
g3012	Cultivated Ivorian	Hybrid	KB11	CNRA - Divo	Ivory Coast	cultivated/improved
g3013	Cultivated Ivorian	G	212	CNRA - Divo	Ivory Coast	cultivated/improved
g3014	Cultivated Ivorian	Hybrid	213	CNRA - Divo	Ivory Coast	cultivated/improved
g3015	Cultivated Ivorian	C	222	CNRA - Divo	Ivory Coast	cultivated/improved
g3016	Cultivated Ivorian	Hybrid	244	CNRA - Divo	Ivory Coast	cultivated/improved
g3017	Cultivated Ivorian	G	410	CNRA - Divo	Ivory Coast	cultivated/improved
g3018	Cultivated Ivorian	G	414	CNRA - Divo	Ivory Coast	cultivated/improved
g3019	Cultivated Ivorian	G	416	CNRA - Divo	Ivory Coast	cultivated/improved
g3020	Cultivated Ivorian	G	Go1	CNRA - Divo	Ivory Coast	cultivated/improved
g3021	Cultivated Ivorian	G	Go2	CNRA - Divo	Ivory Coast	cultivated/improved
g3022	Cultivated Ivorian	G	Po36	CNRA - Divo	Ivory Coast	cultivated/improved
g4001	Ira1	G	02009	CNRA - Divo	Ivory Coast	wild
g4002	Ira1	G	02013	CNRA - Divo	Ivory Coast	wild
g4003	Ira1	G	02017	CNRA - Divo	Ivory Coast	wild
g4004	Ira1	G	02027	CNRA - Divo	Ivory Coast	wild
g4005	Ira1	G	02028	CNRA - Divo	Ivory Coast	wild
g4006	Ira1	G	02037	CNRA - Divo	Ivory Coast	wild
g4007	Ira1	G	02038	CNRA - Divo	Ivory Coast	wild
g4008	Ira1	G	02040	CNRA - Divo	Ivory Coast	wild
g4009	Ira1	G	02044	CNRA - Divo	Ivory Coast	wild
g5001	Ira2	G	02016	CNRA - Divo	Ivory Coast	wild
g5002	Ira2	G	02090	CNRA - Divo	Ivory Coast	wild
g5003	Ira2	G	02093	CNRA - Divo	Ivory Coast	wild
g5004	Ira2	G	02096	CNRA - Divo	Ivory Coast	wild
g5005	Ira2	G	02102	CNRA - Divo	Ivory Coast	wild
g5006	Ira2	G	02104	CNRA - Divo	Ivory Coast	wild

g5007	Ira2	G	02105	CNRA - Divo	Ivory Coast	wild
g5008	Ira2	G	02111	CNRA - Divo	Ivory Coast	wild
g5009	Ira2	G	02129	CNRA - Divo	Ivory Coast	wild
g5010	Ira2	G	02130	CNRA - Divo	Ivory Coast	wild
g5011	Ira2	G	02131	CNRA - Divo	Ivory Coast	wild
g5013	Ira2	G	02151	CNRA - Divo	Ivory Coast	wild
g5014	Ira2	G	02153	CNRA - Divo	Ivory Coast	wild
g5015	Ira2	G	02155	CNRA - Divo	Ivory Coast	wild
g5017	Ira2	G	02158	CNRA - Divo	Ivory Coast	wild
g5018	Ira2	G	02159	CNRA - Divo	Ivory Coast	wild
g5019	Ira2	G	02160	CNRA - Divo	Ivory Coast	wild
g6001	Fourougbankoro	G	02201	CNRA - Divo	Ivory Coast	wild
g6002	Fourougbankoro	G	02202	CNRA - Divo	Ivory Coast	wild
g6003	Fourougbankoro	G	02204	CNRA - Divo	Ivory Coast	wild
g6004	Fourougbankoro	G	02205	CNRA - Divo	Ivory Coast	wild
g6005	Fourougbankoro	G	02207	CNRA - Divo	Ivory Coast	wild
g6006	Fourougbankoro	G	02209	CNRA - Divo	Ivory Coast	wild
g6007	Fourougbankoro	G	02210	CNRA - Divo	Ivory Coast	wild
g6008	Fourougbankoro	G	02212	CNRA - Divo	Ivory Coast	wild
g6009	Fourougbankoro	G	02213	CNRA - Divo	Ivory Coast	wild
g6010	Fourougbankoro	G	02214	CNRA - Divo	Ivory Coast	wild
g6011	Fourougbankoro	G	02215	CNRA - Divo	Ivory Coast	wild
g6012	Fourougbankoro	G	02216	CNRA - Divo	Ivory Coast	wild
g6013	Fourougbankoro	G	02217	CNRA - Divo	Ivory Coast	wild
g6014	Fourougbankoro	G	02218	CNRA - Divo	Ivory Coast	wild
g6015	Fourougbankoro	G	02219	CNRA - Divo	Ivory Coast	wild
g6016	Fourougbankoro	G	02221	CNRA - Divo	Ivory Coast	wild
g6017	Fourougbankoro	G	02224	CNRA - Divo	Ivory Coast	wild
g6018	Fourougbankoro	G	02226	CNRA - Divo	Ivory Coast	wild
g6019	Fourougbankoro	G	02227	CNRA - Divo	Ivory Coast	wild
g6020	Fourougbankoro	G	02228	CNRA - Divo	Ivory Coast	wild
g6021	Fourougbankoro	G	02233	CNRA - Divo	Ivory Coast	wild
g7001	Mouniandougou	G	02930	CNRA - Divo	Ivory Coast	wild
g7002	Mouniandougou	G	02931	CNRA - Divo	Ivory Coast	wild
g7003	Mouniandougou	G	02932	CNRA - Divo	Ivory Coast	wild

g7004	Mouniandougou	G	02933	CNRA - Divo	Ivory Coast	wild
g7006	Mouniandougou	G	02935	CNRA - Divo	Ivory Coast	wild
g7007	Mouniandougou	G	02936	CNRA - Divo	Ivory Coast	wild
g7008	Mouniandougou	G	02937	CNRA - Divo	Ivory Coast	wild
g7009	Mouniandougou	G	02938	CNRA - Divo	Ivory Coast	wild
g7010	Mouniandougou	G	02939	CNRA - Divo	Ivory Coast	wild
g7011	Mouniandougou	G	02940	CNRA - Divo	Ivory Coast	wild
g7012	Mouniandougou	G	02941	CNRA - Divo	Ivory Coast	wild
g7013	Mouniandougou	G	02942	CNRA - Divo	Ivory Coast	wild
g7014	Mouniandougou	G	02943	CNRA - Divo	Ivory Coast	wild
g7015	Mouniandougou	G	02944	CNRA - Divo	Ivory Coast	wild
g7016	Mouniandougou	G	02945	CNRA - Divo	Ivory Coast	wild
g7017	Mouniandougou	G	02946	CNRA - Divo	Ivory Coast	wild
g7018	Mouniandougou	G	02947	CNRA - Divo	Ivory Coast	wild
g7019	Mouniandougou	G	02948	CNRA - Divo	Ivory Coast	wild
g7020	Mouniandougou	G	02949	CNRA - Divo	Ivory Coast	wild
g7021	Mouniandougou	G	02951	CNRA - Divo	Ivory Coast	wild
g7022	Mouniandougou	G	02952	CNRA - Divo	Ivory Coast	wild
g7023	Mouniandougou	G	02953	CNRA - Divo	Ivory Coast	wild
g7024	Mouniandougou	G	02954	CNRA - Divo	Ivory Coast	wild
g7025	Mouniandougou	G	02955	CNRA - Divo	Ivory Coast	wild
g7026	Mouniandougou	G	02956	CNRA - Divo	Ivory Coast	wild
g7027	Mouniandougou	G	02957	CNRA - Divo	Ivory Coast	wild
g7028	Mouniandougou	G	02958	CNRA - Divo	Ivory Coast	wild
g7029	Mouniandougou	G	02960	CNRA - Divo	Ivory Coast	wild
g7030	Mouniandougou	G	02961	CNRA - Divo	Ivory Coast	wild
g7031	Mouniandougou	G	02962	CNRA - Divo	Ivory Coast	wild
g7032	Mouniandougou	G	02963	CNRA - Divo	Ivory Coast	wild
g8001	Sabregue	G	02351	CNRA - Divo	Ivory Coast	Spontaneous
g8002	Sabregue	G	02357	CNRA - Divo	Ivory Coast	Spontaneous
guy1	Guy	Hybrid	Nana1_E1/29/4	CIRAD - French Guyana	Ivory Coast	unknown
guy2	Guy	Hybrid	Nana2-E1/29/7	CIRAD - French Guyana	Ivory Coast	unknown
guy3	Guy	Hybrid	197	CIRAD - French Guyana	Ivory Coast	unknown
guy4	Guy	Hybrid	651	CIRAD - French Guyana	Ivory Coast	unknown
in001	Cultivated Ivorian	Hybrid	269	CNRA - Divo	Ivory Coast	cultivated/improved

in002	Cultivated Ivorian	SG2	345	CNRA - Divo	Ivory Coast	cultivated/improved
in003	Cultivated Ivorian	Hybrid	149	CNRA - Divo	Ivory Coast	cultivated/improved
in004	Cultivated Ivorian	SG2	028	CNRA - Divo	Ivory Coast	cultivated/improved
ue005	Erect	SG2	Not available	Naro-CORI	Uganda	cultivated/improved
ue006	Erect	SG2	Not available	Naro-CORI	Uganda	cultivated/improved
ue007	Erect	SG2	Not available	Naro-CORI	Uganda	cultivated/improved
ue011	Erect	SG2	Not available	Naro-CORI	Uganda	cultivated/improved
un002	Nganda	SG2	Not available	Naro-CORI	Uganda	cultivated/improved
un004	Nganda	SG2	Not available	Naro-CORI	Uganda	cultivated/improved
un008	Nganda	SG2	Not available	Naro-CORI	Uganda	cultivated/improved
un016	Nganda	SG2	Not available	Naro-CORI	Uganda	cultivated/improved
un020	Nganda	SG2	Not available	Naro-CORI	Uganda	cultivated/improved
un025	Nganda	SG2	Not available	Naro-CORI	Uganda	cultivated/improved
un031	Nganda	SG2	Not available	Naro-CORI	Uganda	cultivated/improved
un037	Nganda	SG2	Not available	Naro-CORI	Uganda	cultivated/improved
uw025	UW	UW	Not in collection	Naro-CORI	Uganda	wild
uw100	UW	UW	Not in collection	Naro-CORI	Uganda	wild
uw219	UW	UW	Not in collection	Naro-CORI	Uganda	wild

Table S2: Pairwise F_{st} for the diverse levels of structure investigation

1st level (2 groups)

	Guinean
Congolese	0.250

2nd level (4 groups)

	Nana	Other Congo	Pelezi
Other Congo	0.306		
Pelezi	0.450	0.354	
Other Guinean	0.366	0.313	0.227

3rd level (6 groups)

	SG2	Libenge	Nana	Niaouli	Pelezi
Libenge	0.241				
Nana	0.336	0.433			
Niaouli	0.303	0.431	0.422		
Pelezi	0.401	0.497	0.455	0.503	
Other Guinean	0.342	0.424	0.374	0.419	0.226

Other Guinean

	Pine	Cultivated Ivorian	Ira1	Ira2	Fourougbankoro
Cultivated Ivorian	0.112				
Ira1	0.193	0.071			
Ira2	0.231	0.131	0.157		
Fourougbankoro	0.153	0.065	0.121	0.135	
Mouniandougou	0.131	0.111	0.182	0.241	0.133

6 groups with population for Other Guinean

	SG2	B	Nana	SG1	Mouniandougou	Fourougbankoro	Ira2	Ira1	Pine	Pelezi

B	0.241									
Nana	0.336	0.433								
SG1	0.303	0.431	0.422							
Mouniandougou	0.402	0.501	0.431	0.505						
Fourougbankoro	0.349	0.449	0.387	0.442	0.133					
Ira2	0.372	0.478	0.452	0.487	0.240	0.135				
Ira1	0.367	0.490	0.430	0.479	0.180	0.122	0.158			
Pine	0.350	0.465	0.387	0.458	0.129	0.152	0.230	0.192		
Pelezi	0.401	0.497	0.455	0.503	0.352	0.258	0.339	0.307	0.239	
Cultivated Ivorian	0.320	0.436	0.370	0.411	0.111	0.065	0.131	0.072	0.112	0.246

All Fst values significant at a 5% threshold

Table S3: F_{st}-based AMOVAs and derived *F-statistics* on different levels of structure investigation

		Source of variation				Related *F-statistics*				
Level	**AMOVA Design**	**Among groups**	**Among populations or among populations within groups**	**Among individuals within populations**	**Within individuals**	***Fis***	***Fit***	***Fst***	***Fsc***	***Fct***
1st level (2 clusters)	2 populations (Guinean and Congolese) corresponding to the 1st level of Structure	*NA*	24.94	30.74	44.33	0.40949	0.55675	0.24938	*NA*	*NA*
2nd level (4 clusters)	2 groups (Guinean and Congolese), 2 populations per group: Pelezi and Other Guinean for the Guinean group, Nana and Other Congolese for the Congolese group	9.63ns	24.72	21.32	44.33	0.32479	0.55671	*NA*	0.27349	0.09632ns
3rd level (6 clusters)	2 groups (Guinean and Congolese), 2 populations for Guinean: Pelezi and Other Guinean, 4 populations for Congolese: Nana, Libenge and SG2	11.75ns	27.52	17.11	43.62	0.28174	0.56381		0.31185	0.11751ns
Other Guinean (6 clusters)	1 group (Other Guinean), 6 populations	*NA*	14.18	25.46	60.36	0.29666	0.39642	0.14184		
3rd level with population information for Other Guinean (11 clusters)	2 groups (Congolese and Guinean), 4 populations for Congolese and 7 populations for Guinean	20.1	22.02	13.87	44.02	0.23957	0.55985		0.27558	0.20098

Table S4: F_{is} per population on different levels of structure investigation

1st level (2 clusters)	Congolese	0.414
	Guinean	0.405
2nd level (4 clusters)	Nana	0.210
	Other Congolese	0.339
	Pelezi	0.197
	Other Guinean	0.380
3rd level (6 clusters)	Nana	0.205
	Libenge	0.124
	Niaouli	0.147
	SG2	0.253
	Pelezi	0.193
	Other Guinean	0.380
Other Guinean (6 clusters)	Pine	0.374
	Cultivated Ivorian	0.325
	Ira1	0.226
	Ira2	0.213
	Fourougbankoro	0.273
	Mouniandougou	0.289
3rd level with population information for Other Guinean (11 clusters)	SG2	0.253
	B	0.124
	Nana	0.205
	SG1	0.147
	Mouniandougou	0.287
	Fourougbankoro	0.273
	Ira2	0.213
	Ira1	0.244
	Pine	0.373
	Pelezi	0.193
	Cultivated Ivorian	0.325

Article soumis à Genome

A.3.5 : Différenciation génétique de populations sauvages et cultivées : diversité de Coffea canephora en Ouganda

Genetic differentiation of wild and cultivated populations:

Diversity of *Coffea canephora* in Uganda.

P. Musoli [1], P. Cubry[2], P. Aluka[1], C. Billot[2], M. Dufour[2], F. De Bellis[2], D. Pot[2], D. Bieysse[2], A. Charrier[3], T. Leroy[1*]

[1] Coffee Research Institute, P.O. Box 185, Mukono, Uganda.

[2] CIRAD, UMR DAP, TA-A96/03, Avenue Agropolis, 34398 Montpellier CEDEX 5, France;

[3] Supagro, 2 Place Viala, Montpellier, F-34060 France

[*] Corresponding author: Thierry Leroy, CIRAD TA A 96/03, Avenue Agropolis, 34398 Montpellier Cedex 5, France.

Fax number (33) 4-67-61-57-93

Email: thierry.leroy@cirad.fr

Abstract

Wild, feral, and cultivated populations of *Coffea canephora* coffee trees from Uganda were evaluated using neutral markers. 196 Ugandan *C. canephora* genotypes from 14 sites were analysed using 24 microsatellite markers. Basic diversity, dissimilarity and genetic distances between individuals, genetic differentiation between populations, and population structures were analysed. The wildness of genotypes sampled from Ugandan primary forests was confirmed. Four groups of genotypes were identified within Ugandan genotypes, discriminating between feral, cultivated and two wild origins. Ugandan wild populations showed a clear pattern of isolation by distance. We observe and discuss the relationships between those populations. We compared the Ugandan populations with known genetic diversity groups within the species using 18 markers. An analysis of genetic distances was performed using genetic differentiation parameters. *Coffea canephora* of Ugandan origin was found to be different from known diversity groups, implying that it forms another diversity group within the species. Hypotheses on the differentiation of *C. canephora* since the Last Glacial Maximum are discussed.

Résumé

La variabilité génétique des populations sauvages et cultivées du caféier *Coffea canephora* provenant d'Ouganda a été analysée avec des marqueurs neutres. 196 génotypes ougandais provenant de 14 sites ont été analysés avec 24 marqueurs microsatellites. La diversité des populations, les dissimilarités et les distances génétiques entre les individus, la différentiation génétique entre les populations et la structure des populations ont été analysées. Le caractère sauvage des génotypes prospectés dans les forêts primaires d'Ouganda est confirmé. Quatre groupes de génotypes sont identifiés dans les origines ougandaises, discriminant les origines sauvages, cultivées et férales. Un isolement par la distance est mis en évidence dans les populations sauvages d'Ouganda. Par ailleurs, nous discutons des relations entre les génotypes de *C. canephora* d'origine cultivée, férale ou sauvage en Ouganda. 18 marqueurs ont été utilisés pour comparer les populations de caféiers ougandais à celles des autres groupes de diversité déjà identifiés dans l'espèce. Une analyse des distances génétiques a été conduite en utilisant des paramètres de différentiation génétique. Les génotypes de *Coffea canephora* d'Ouganda s'avèrent être différents de ceux des autres groupes préalablement identifiés, ils forment donc un nouveau groupe de diversité de l'espèce. Nous discutons d'hypothèses sur la différentiation de l'espèce *C. canephora* depuis le dernier maximum glaciaire.

Introduction

In Africa, at least three centres of diversity have been identified and were likely refugia during Last Glacial Maximum (LGM), namely the Guinean region, the central continental region and the coastal Atlantic region (Hamilton 1976; White 1979). In addition, some authors contend that there is another "East central Africa" refugium located in the interlacustrine Highlands bordering the Albertine rift and including Western Uganda (Jolly et al. 1997; Anhuf et al. 2006).

The *Coffea* genus (*Rubiaceae*) is endemic to the tropical forests of Africa, from Guinea to eastern Africa and Madagascar. This genus includes about 100 species. Coffee is an extremely important crop with nearly 7 million tons of green beans produced yearly in about 80 tropical countries. In terms of economic importance on international markets, it is second to oil, earning more than 9,000 million US $ per year. There are two main types of cultivated coffee, Arabica and Robusta. Arabica is produced from *Coffea arabica*, which grows in highlands, while Robusta is produced from *Coffea canephora*, which is cultivated at low to medium altitudes (35% of global production). *Coffea canephora* Pierre ex Frohener is self-incompatible and diploid while *Coffea arabica* L. is tetraploid and self compatible (Charrier and Berthaud 1985). *Coffea canephora* is indigenous to the tropical African forest, stretching from West Africa through Cameroon, Central African Republic (CAR), Congo, the Democratic Republic of Congo (DRC), Uganda, and northern Tanzania up to northern Angola. The *C. canephora* populations are generally small disconnected populations with a small number of mother-trees and few offspring scattered over areas smaller than one ha.

The genetic diversity of *C. canephora* was first described by Berthaud in 1986 using isozymes and studying wild and cultivated genotypes from western and central Africa. He identified two diversity groups, a Congolese group, which comprised genotypes from CAR and Cameroon, and a Guinean group, which consisted of genotypes of wild origin from Ivory Coast. Montagnon et al. (1992), also using isozymes, proposed a substructure in the Congolese group, with two sub-divisions SG1 & SG2. Dussert et al. (2003), using RFLP (Restriction Fragment Length Polymorphism) molecular markers placed *C. canephora* genotypes of cultivated and wild origins into five diversity groups, adding two groups, B and C, to the Congolese group. Recent studies using microsatellites on *C. canephora* among other coffee species (Poncet et al. 2004) and on *C. canephora* alone (Cubry et al. 2005) confirmed the latter structure.

According to the literature (Thomas, 1935), *Coffea canephora* was found wild and domesticated in Uganda and northern Tanzania long before colonization. It was developed as a plantation crop from the end of 19th century by introducing seeds, mainly from the Belgian Congo (Thomas, 1947). Due to self incompatibility of the species and the proximity of cultivated fields to the forest, one expects to find an admixture between various origins, including local wild genotypes, both in the wild and in coffee plantations. However, no studies have examined the genetic structure of Ugandan wild and cultivated populations. Such an analysis is important to determine the diversity and specificity of *C. canephora* genotypes in Uganda, and their potential use in coffee breeding.

This paper presents the results of a study on Ugandan wild, feral, and cultivated *C. canephora* genotypes, and compares them to genotypes from known diversity groups. In this investigation, we addressed three questions, i) What is the genetic diversity and structure of *C. canephora* within Uganda? ii) Where are Ugandan genotypes located in global *C. canephora* diversity? iii) How did the differentiation of these populations occur in the coffee diversification process in Africa?

Materials and methods

Population sampling

For the purpose of studying Ugandan populations and testing relationships between cultivated and wild populations, hierarchical sampling was performed, covering three regions and two cultivated types.

Wild and feral compartment

The first wild region, Itwara forest, covers an area of 100 km². Five sites separated by distances ranging from 0.6 to 10 km were chosen. At each site, seven to 18 individuals were sampled, for a total number of 55 from the region. The second wild region, Kibale forest, covers 500 km². Four to 30 individuals per site were sampled at four sites separated by a distance of seven to 19 km. 54 individuals were sampled in that region. In both the Kibale and Itwara regions, sampled trees were healthy-looking and estimated to be 40 to 70 years old. This sampling strategy was chosen according to the spatial dispersal of wild coffee populations: few adult trees and volunteers scattered all around the mother-trees. Sampling sites in these regions were located throughout the forests and at one to five kilometres from the edges of the forest. Kibale and Itwara are primary forests considered to be a natural home for wild *C. canephora* (Maitland 1926; Thomas 1944). The two forests once formed a continuum but in the early 1900s areas in between were allocated to human settlement, and they now lay 30 km apart.

The third region, Kalangala, consisted of five sites separated by at least 10 km on two islands in Lake Victoria. They were located 0.5 to two kilometres from the edge of the islands. Three to 12 healthy looking trees over 40 years old were sampled per site, for a total number of 35 individuals sampled in this region. The environment consists of a secondary forest, regenerated from former cultivated areas, and coffee populations are thus considered as feral (Thomas 1935).

Cultivated compartment

Ugandan coffee farmers consider two main cultivated coffee types. Nganda has a spreading growth habit and develops many secondary stems from suckers growing on the main stems. Erect has upright main stems. It is often believed that the latter type was introduced from the Congo Basin, whereas the former may be the wild type, introduced from local forests (Thomas 1935). However, this hypothesis is not substantiated. Among cultivated genotypes, individuals of both the Nganda (31 individuals) and Erect (21 individuals) types were sampled from the Kawanda collection in Uganda. Individuals of each type were selected on the basis of historical records in addition to their phenotypic appearance. In our study, Nganda and Erect genotypes were considered as separate groups equivalent to regions.

Sampling outside Uganda

In order to assign Ugandan *C. canephora* to a diversity group within the species, we selected controls from the Guinean and Congolese groups, i.e. SG1, SG2, B and C (Figure 1). Based on the results of a previous study (Cubry et al. 2005), a subset of 36 genotypes encompassing known *C. canephora* diversity was selected. This subset included the SG1 (2 genotypes), B (7 genotypes), C (5 genotypes), SG2 (7 genotypes, 2 sites) and Guinean (15 genotypes, 2 sites) groups. Guinean and Congolese populations SG1, B and C genotypes are of wild origin while SG2 genotypes are cultivated. Figure 2 presents the plant material analysed in our study.

DNA extraction

Genomic DNA was extracted from ground frozen leaves followed by a MATAB buffer based extraction method adapted from a procedure used for cocoa (Risterrucci et al. 2000). The extracts were purified using the solution-based Promega Wizard Genomic DNA Purification Kit. The concentration of the extracts was estimated with a spectrophotometer and standardized at 0.5 ng/µl final work solution.

SSR genotyping

Twenty-four polymorphic microsatellite markers (SSR) mapped on the *C. canephora* genome were used to genotype the 196 individuals from Uganda (Table 1). Markers DL013

and DL025 were previously designed from a BAC library developed for studying sugar metabolism in coffee (Leroy et al. 2005). The other sets came from a SSR enriched library of *C. canephora* clone 126 (Dufour et al. 2002) and from enriched libraries of *C. arabica* var. Caturra (Combes et al. 2000; Rovelli et al. 2000). These markers have been assessed for their amplification in *C. canephora* and related species (Poncet et al. 2007; Cubry et al. 2008). All the microsatellites were previously mapped on our genetic map (T. Leroy, personal communication, 2006). In order to avoid any redundancy in diversity information, the chosen markers covered nine of the eleven linkage groups (n=11), and were located at a distance of at least 50 cM from each other.

A subset of 18 from the 24 SSR markers used in a previous study (Cubry et al. 2005) was considered for the comparative analyses between Ugandan genotypes and other diversity groups.

PCR amplification and visualization of the microsatellites

PCR reactions were performed in 10 µl, containing 2.5 ng of DNA, 1 mM Tris-HCl, 5 mM KCl, 2 mM MgCl2, 200 µM dNTP, 0.10 µM of reverse primer, 0.08 µM of forward primer tailed with M13 sequence, 0.10 µM of infrared fluorescently-labelled M13 primer and 0.1 U of Taq DNA polymerase. PCR amplifications were run in an Eppendorf Ep384 thermocycler. The amplification programme consisted of an initial denaturation cycle of 4 min at 94°, followed by 10 cycles of "Touch-Down" (45 sec at 94°, 1 min at 60° to 55° decreasing by 0.5° per cycle, and 1.5 min at 72°), 25 cycles (45 sec at 94°, 1 min at 55°, and 1.5 min at 72°) and ended with a final elongation step at 72° for 5 minutes.

Fluorescently-labelled PCR products were analysed by electrophoresis migration on a LiCor® 4300 automated sequencer with a 6.5% acrylamide gel. The gel images were retrieved and annotated with the manufacturer's program, SAGA® GT Generation Two. Allele sizes were evaluated on the basis of allelic controls previously defined by Cubry et al. (2005). Data matrix was used for further analyses. All genotyping experiments were performed on the Genopole Genotyping Platform, Montpellier, Languedoc-Roussillon, France.

Within Uganda analysis – 24 marker set

Basic diversity analyses

Summary statistics, including the number of alleles, along with observed and expected heterozygosity (gene diversity), were calculated for all groups of genotypes using PowerMarker (Liu and Muse 2005). Significance was tested using a bootstrap procedure (1,000 repetitions) to determine standard deviation and a 95% confidence interval.

CONVERT software (Glaubitz 2004) was used to point out private alleles in all regions and sites. This software was also used to format the data for other software analyses (see below), including Fstat (Goudet 2001), Arlequin (Excoffier et al. 2005) and Structure (Pritchard et al. 2000).

Dissimilarity analysis

The dissimilarity matrix between Ugandan individuals using 24 SSR markers, based on a simple matching index, was computed with DARwin 5 (Perrier et al. 2003) taking into account missing data. A tree representation of the dissimilarities was obtained with the same software and built with the weighted neighbour joining method (Saitou and Nei 1987). The robustness of the nodes was assessed by 1,000 bootstraps.

Fixation and genetic differentiation analyses

On the different sampling levels within Uganda, Hardy-Weinberg equilibrium (HWE) was tested using fixation indices within sites, regions or country (F_{IS}). Genetic differentiations between site or region levels were estimated with F_{ST} for all regions. All *F*-statistics were estimated with Fstat software (Goudet 2001) using estimations proposed by Weir and Cockerham (1984), since Nei's *F*-statistics (Nei 1973) are less convenient when sample sizes vary substantially. Although SSR markers might not follow assumptions needed for *F*-statistics, especially in terms of mutation rate, these statistics were preferred to R_{ST} (Slatkin 1995), since in our case (small number of individuals in some cases, and small number of loci) they appeared to be more robust (Gagiotti et al. 1999). The significance of all statistics was tested with 1,000 permutations of the genotypes (Goudet et al. 1996).

Isolation by distance study

In order to test whether genetic differentiation followed a pattern of isolation by distance, we performed a Mantel test using GenAlEx software (Peakall and Smouse 2006). We applied the data transformation proposed by Rousset (1997), to perform a Mantel test between the logarithm of the geographical distance and a genetic distance coefficient calculated as $F_{ST} / (1-F_{ST})$. This test was used to calculate correlations within Ugandan wild populations only (Kibale and Itwara) and determine the possible relationships at short geographical distances, on a forest scale. Geographical distances between populations were computed using GPS coordinates for each site.

Molecular variance and population structure analysis

The level on which major genetic differentiation occured was estimated by hierarchical Analyses of MOlecular VAriance (AMOVA) using Arlequin 3.0 software (Excoffier et al. 2005). Analyses were performed on region, population and individual levels,

with an accepted level of missing data of 0.2 and tested with 1,000 permutations. Molecular distance was computed using the number of different alleles.

Finally, the fine structure and relationships of Ugandan populations were analysed with STRUCTURE 2.1 software (Pritchard et al. 2000). STRUCTURE was used to infer the population structure within Uganda, assign individuals to groups, and discover more about admixtures between defined populations. A 50,000 burn-in period and 100,000 MCMC iterations were used for each run. We chose to take the default parameters with no prior information about the populations' origin in order to reallocate our individuals to groups. Simulations were made for a number of groups, K varying from K = 1 to K = 20 with five repetitions for each value of K. Transformation of the resulting data, using a method proposed by Evanno et al. (2005) that took into account the rate of change in log probability of data between successive K values, enhanced the detection of optimum K values.

Ugandan diversity in the whole C. canephora species – 18 marker set

Relative place of Ugandan genotypes

In order to study the place of Ugandan populations within the whole diversity of the species, a pairwise genetic distance matrix of F_{ST} was built using 18 markers and the previously described procedure using Fstat. A WPGMA tree was computed using F_{ST} linearized values ($F_{ST}/(1-F_{ST})$) with DARwin 5 software for all regions including Uganda.

Molecular variance and population structure analysis

We performed an AMOVA as described for Ugandan genotypes on a species level, considering groups SG1, SG2, B, C, G and Uganda. See the AMOVA design in Table 6.

Isolation by distance analysis on a species level

For all regions including Uganda, a Mantel test was performed in order to find possible relationships on the overall level throughout Africa. The geographical distances between diversity regions for overall analysis were estimated by comparing collection sites for groups B and C with Ugandan wild populations. An average geographical centre of the group was used for regions SG1 and SG2, since no GPS or precise geographical data were available for those regions.

Results

Ugandan level

Basic diversity statistics and diversity tree

Table 2 presents a summary of basic diversity statistics. Among the Ugandan samples, the Kibale and Itwara regions gave lower values for observed heterozygosity and gene diversity than cultivated populations and the Kalangala region. The Itwara region had higher

values for observed heterozygosity and gene diversity than the Kibale region. For the three cultivated and feral Ugandan populations, we observed heterozygosity and gene diversity values close to the values of the cultivated SG2 group (Cubry et al., 2005). Kibale and Iwara had heterozygosity values close to those obtained for other known wild populations: Guinean, Congolese B and C (Cubry et al., 2005). Among the Ugandan regions, wild forest regions and Kalangala had the largest number of private alleles. It was found that fewer than 10% of the individuals harboured those private alleles, except for one locus in the Itwara region. The number of private alleles was much larger in the Itwara than in the Kibale populations.

The diversity tree for the 196 individuals from Uganda is presented in Figure 3. Four groups of genotypes were identified, discriminating between Kibale, Itwara, and Kalangala, and a group that included Erect and Nganda individuals. Three Nganda genotypes and one Kalangala genotype were located close to the Kibale genotypes.

Fixation indices and genetic differentiation

Among the Ugandan individuals, all the F_{IS} values were significantly different from zero, except for two sites displaying a small number of individuals (7), one in Kibale and one in Itwara (Figure 4). As far as wild populations were concerned, since *C. canephora* is a strictly allogamous species, this may be related to the Wahlund effect, i.e. some structure within sites or regions. This structure would seem to be related to limited pollen flow and or seed dispersal and thus to relatedness between individuals. This may also be a result of the founder effect, since most wild populations are normally in patches comprising a small number of mother-trees (Berthaud 1986). For cultivated genotypes such as Nganda or Erect, it may involve bias related to the nature of the groups which include genotypes from different origins.

In order to assess population differentiation, F_{ST} coefficients were estimated between sites for each region and between regions for Uganda. When considering the five Ugandan regions differentiation was significant, with a F_{ST} of 0.16 (Figure 4). Within Ugandan regions, the results were different for between-site differentiation for the Kibale, Itwara and Kalanga regions. The Kibale region exhibited a highly significant between-site F_{ST} coefficient (F_{ST} =0.17), meaning that this region was highly differentiated into sites. This was probably due to large distances between sites and thereby fewer genetic exchanges. These results tallied with other analyses on populations from the same forest (Nyakaana 2007). On the other hand, F_{ST} coefficients were not significant for the Itwara forest (0.017) and Kalangala region (0.039), meaning that sites sampled inside those regions were not genetically different. This was probably due to shorter distances between sites and thereby more genetic exchanges when

compared to sites in the Kibale region. For Kalangala, successive introductions of cultivated genotypes might also explain these results.

Table 3 presents F_{ST} values among *C. canephora* populations within Uganda. The values observed between the two cultivated Ugandan regions (Nganda and Erect) and the feral region were low, ranging from 0.08 to 0.14, confirming their common genetic background. F_{ST} values for relationships between cultivated/feral and wild populations were much higher, ranging from 0.15 to 0.36. Pairwise F_{ST} values between sites were very low for the Kalangala region (0.00 to 0.11) and Itwara (0.00 to 0.05), but quite high (0.11 to 0.35) between sites in the Kibale region, confirming the existence of substantial differentiation between sites in Kibale.

Isolation by distance within the Ugandan wild compartment

Ugandan wild populations showed a clear pattern of isolation by distance (P_{Mantel}=0.015, Table 5). The Itwara population showed a pattern of isolation by distance (P_{Mantel}=0.04, data not shown), whereas the Kibale population did not show that pattern.

Molecular variance and population structure

The results of the Analysis of MOlecular VAriance (AMOVA) showing the genetic level on which the main differentiations occurred are presented in Table 6. Taking into consideration only Ugandan samples with 5 groups, 2 wild (Kibale and Itwara) and 3 cultivated/feral (Kalangala, Erect, Nganda), the percentage of variation explained by regions was not high (13.5%), in comparison to variation between individuals within regions (19.9%) and within individuals (66.7%).

The fine structure and relationship results for the Ugandan populations analysed by Structure indicated four populations (Figure 6). The Nganda and Erect populations constituted a single group. Wild populations, Kibale and Itwara, were clearly identified as different groups, with some degree of admixture. We also observed some gene flows from the Kibale region to the Nganda-Erect group. Finally, the Kalangala islands were identified as a specific group, with low admixture with other populations.

Ugandan genotypes in species diversity

Genetic differentiation analysis

Table 4 presents F_{ST} coefficients between all the *C. canephora* regions in Africa. F_{ST} values for diversity regions excluding Uganda were over 0.20, indicating high differentiation between regions. Considering Uganda in relation to other regions, it is important to note that pairwise F_{ST} values were always lower for Ugandan regions when relating to SG2 than when

relating to other regions. This implied that Ugandan *C. canephora* are genetically more related to the SG2 region. F_{ST} values obtained between wild Ugandan regions and the other Congolese and Guinean diversity groups were much greater than those obtained between regions within Uganda (0.28 to 0.59). A F_{ST} value of 0.10 was observed between the Kibale and Itwara regions.

The WPGMA tree constructed with the linearized F_{ST} matrix for all regions is presented in Figure 5. The Itwara and Kibale regions, which were closely related to each other, could be considered as one original diversity group. Other Ugandan regions (cultivated and feral) were close to the SG2 group, confirming a common genetic background with these Congolese genotypes. Based on this tree, 6 diversity groups were identified within *C. canephora*: Guinean, C, SG1, B, Itwara and Kibale as one "Ugandan" diversity group, and the SG2 group including cultivated and feral genotypes from Uganda.

AMOVA analysis on a species level

Table 6 presents AMOVA results. When six groups were considered, wild from Uganda, Guinean, Congolese B, C, SG1 and SG2, variation was mostly partitioned between groups (36.2%) and among individuals within groups (23.4%). This result indicates that the major differentiation occurred on a group level, confirming the structure of *C. canephora* diversity in 6 groups.

When considering cultivated and feral populations from Ugandan (Nganda, Erect and Kalangala) as an additional group in the study, defining 7 groups, the results indicated that a high percentage of variation was explained by the groups defined (20.3%). A small percentage of variation was explained by the 5 regions within the Ugandan group (6.9%) and more than 70% of variation was explained by individuals among regions and within individuals. This result indicated low differentiation within Uganda, compared to that observed for the 6 diversity groups.

When considering all the diversity regions throughout Africa, non-significant correlations were found between geographical and genetic distances (Table 5).

Discussion

The analyses carried out on the *Coffea canephora* genotypes from Uganda revealed their originality and diversity. Four groups of genotypes were identified within Ugandan genotypes, discriminating between feral, cultivated and two wild origins. We compared Ugandan populations with known genetic diversity groups from western and central Africa within the species using genetic differentiation parameters. *Coffea canephora* trees of wild

origin in Uganda were found to be different from known diversity groups, while cultivated and feral genotypes were identified as being close to one Congolese diversity group (SG2).

Refuge zones during the LGM; differentiation process in Africa

The climatic and vegetation changes occurring in Africa between the Cenozoic period and the present day have been described (Maley 1996). The Guinean-Congolese region consists of three centres of diversity related to the last glaciations and refugia (White 1979): the Guinean region, the central continental region, and the coastal Atlantic region, the last one being the most valuable in terms of genetic diversity (Jolly et al. 1997). Our results obtained on *C. canephora* confirmed the importance of these three zones, corresponding to the Guinean, SG2 and SG1 regions respectively, which were refugium zones during the last glaciations for forest trees (Maley 1996; Adams and Faure 1997). In Uganda, some mountain forests remained during the last glaciations (Jolly et al. 1997) and the Ugandan region may correspond to the "East central Africa" refugia (Anhuf et al. 2006). The high genetic differentiation of Ugandan wild genotypes tends to confirm the existence of this refugium zone for *C. canephora*. Coffee's geographical spread from refugium zones occurred by seeds and pollen dispersal. Pollen dispersal by insects may be possible for up to two or three km in forest areas. Dispersal of seeds by animals, especially mammals, may be possible for up to five or 10 km (Berthaud, 1986). In Uganda, genotypes may have migrated from the Western Albertine rift refugia to western Uganda, where they were identified at the beginning of the 20th century (Maitland, 1926). In Ivory Coast, as in Uganda, the coffee populations surveyed were isolated from the others by distances over 10 km. These personal observations mean that either short-distance dissemination or rare dissemination events over long distances can occur for coffee. Systematic sampling of all forests throughout the tropical African forest has not yet been possible. Some large zones between refugium zones in DRC, Gabon, Angola and the Central African Republic would provide more information about genetic diversity and relationships between regions.

Differentiation and structure of populations

Our study clearly discriminated between the different origins, confirming genetic diversity groups described in earlier studies using other molecular markers (Dussert et al. 2003) and revealed the specificity and great diversity of *C. canephora* from Uganda (Figure 2).

Among the Ugandan populations, wild and cultivated *C. canephora* were clearly identified and discriminated between (Figure 2 and Figure 5). Both wild populations exhibited contrasting results. While Kibale was sampled over a larger area than Itwara, and although the

same number of individuals was sampled, observed heterozygosity, gene diversity and allelic content were lower in Kibale. This could be a consequence of sampling that was not equally balanced between sites, with a large share of the individuals sampled at one site for this forest (site 2), while the others were under-sampled. In addition, genetic differentiation between sites was higher than in Itwara, which may be related to the pattern of effective dispersion of pollen and seeds, with sites being chosen at distances greater than the effective neighbourhood size. The level of introgression observed for Itwara region alleles into Kibale and for Kibale alleles into Nganda and Erect (Figure 6) firstly suggested that there was gene flow between Itwara and Kibale, and secondly that the Kibale samples were used as a source for cultivated populations.

We observed substantial genetic diversity within the Kalangala islands. Those feral populations displayed a large number of alleles and private alleles. A strong structure was exhibited within the sites sampled on two islands, while no differentiation between sites was observed. This origin is a source of diversity for coffee in Uganda, and has been reported as a source of material for resistance to Coffee Wilt Disease (Musoli, 2007).

Nganda and Erect displayed the same genetic background (Figures 2 and 6), probably because their phenotypic differences were not reflected by neutral genetic differentiation, as has already been observed in tomato (Miller and Tanksley 1990). Similarities between the Nganda and Erect origins can be explained by their cultivation history. Both origins have been cultivated closely in mixtures for over three generations and open pollinated seedlings were used for planting new generations in this self-incompatible species. These cultivated trees were close to the SG2 diversity group.

The AMOVA indicated a low percentage of variation explained by *C. canephora* regions from different Ugandan sources. These results showed that all Ugandan populations had a common genetic background and suggested that genetic exchange could have occurred between these populations and/or that they could be derivates from an original unique population.

On a species level, analyses of molecular variances confirmed high differentiation between regions. We confirmed a structure with six diversity groups, the five already identified and a wild Uganda should be considered for global *C. canephora* diversity. Within Congolese regions B and C, where only one site was surveyed, we might not be very accurate regarding global diversity. Wider and systematic sampling from these forests and others, including relict forests near sampled primary forest, for each Ugandan and the Congolese regions B and C would give more precise indications.

History and relationships of wild/cultivated

Admixtures observed between the Kibale and Itwara forests confirmed their previous connection, but it concerned about 10% of the population, although the distance was less than 30 kilometres and the split occurred 100 years ago. Thus that distance was sufficient to prevent gene flow between wild *C. canephora* genotypes. Relationships between the Kibale forest and Nganda-Erect population highlight the complexity of coffee genetic diversity in Uganda. As in most other African countries, cultivated coffee appeared as a result of natural crosses between wild materials and introduced genotypes from other regions or countries, leading to a complete mixture of all genotypes (Montagnon et al. 1998). Historical papers from Uganda (Thomas 1944; Leakey 1970) revealed the difficulty in clearly tracing the source of cultivated trees and clearly confirming the wild or cultivated origins of *C. canephora* populations in Uganda. As reported in previous papers, *C. canephora* was being cultivated and traded in Uganda for use in traditional ceremonies earlier than 1800, before commercial cultivation started (Maitland, 1926). The source of planting materials for those early planting has not been clear and Thomas (1935, 1944) suggested the wild origin of Nganda genotypes. Genetic relationships observed between Kibale and Nganda-Erect populations support that hypothesis and indicate Kibale as a likely source of early Nganda establishments in Uganda, as suggested by Thomas (1935). The selection processes and mixed cultivation of Nganda and Erect genotypes in research station fields with introductions from the Congo basin also explain the separate evolutions of these wild and cultivated populations, and the genetic relationships between cultivated Ugandan materials and the SG2 group.

The specificity of the Kalangala region is particularly interesting in this study. Kalangala is one of the first centres of *C. canephora* cultivation in Uganda. While an insularity syndrome was expected (isolation of populations), different sources of materials were identified. An introduction from the defunct Kasai forest is documented (Thomas, 1935). That forest does not exist anymore and was therefore not sampled. However, at the beginning of the 20th Century, farmers were provided with seeds from parents of Nganda phenotypes, probably of wild origin. After the first period of cultivation, coffee plantations were abandoned for 15 years and when cultivation resumed farmers were supplied with seeds derived from Erect genotypes (Thomas 1947). The same author explains that most of the Erect genotypes introduced into Kalangala from the mainland in the 1920s were not adapted to island conditions (low soil pH). Thus, the old trees we surveyed could have mainly originated from the Kasai forest. This history of coffee growing in this region, and isolation

on islands, could explain the specificities observed in the Kalangala *C. canephora* population. It suggests the Kalangala *C. canephora* population is mainly a mixture of populations from forest and from cultivated types. Finally, introgression from Erect/Nganda appeared to be of low incidence (less than 5% in STRUCTURE analysis) while introgression from the Kasai forest could not be quantified.

From our studies it can be deduced firstly that cultivated Ugandan coffees originated from wild material in forests covering parts of the country near the Ruwenzori Mountain and areas near Lake Victoria, including the Kibale and Itwara regions. Genotypes from these sources have been cultivated in admixtures and perhaps with introduced genotypes from the SG2 region, leading to the Nganda and Erect cultivated types. Compared to most diversity groups, Ugandan *C. canephora* is highly diverse and rich in alleles and Uganda can therefore be considered as another centre of diversity.

Consequences for diversity management and breeding

For some crops like coffee, the breeding strategy is already based on hybrid vigour and complementary characteristics obtained by crossing genotypes from distinct genetic and geographical groups in a recurrent breeding strategy (Leroy et al., 1993). Ugandan genotypes could be integrated into the recurrent selection scheme as a new diversity group to be hybridized with Guinean and Congolese genotypes. These results on a tropical tree crop in Africa would lead to the following consequences for diversity management and breeding: (i) complete the genetic diversity of the species by systematic surveys in the putative refugium zones/centres of diversity; (ii) analyse the genetic structure of those genotypes; (iii) propose breeding strategies using that structure to improve productivity, product quality and resistance to diseases and insects.

Acknowledgments

This project was carried out in connection with European -INCO project ICA4-CT-2001-10006 on genetic resistance to coffee wilt in Uganda, and with financial support from the United States Department of Agriculture within the framework of a project of the USDA-International Centre for Research in Agro Forestry, project 58-4001-3-F157 on coffee quality and markers in East Africa. Samples from diversity regions outside Uganda were kindly provided by Centre National de Recherche Agronomique in Ivory Coast. We are grateful to the Uganda Wild Life Authority for allowing us to take samples in the Kibale forest.

References

Adams, J.M., and Faure, H. 1997. Review and Atlas of Paleovegetation: Preliminary land ecosystem maps of the world since Last Glacial Maximum. TN, USA, Oak Ridge National Laboratory, available at http://esd.ornl.gov/projects/gen/adams4.html.

Anhuf, D., Ledru, M.P., Behling, H., Da Cruz, F.W., Cordeiro, R.C., Van der Hammen, T., Karmann, I., Marengo, J.A., De Oliveira, P.E., Pessenda, L., Siffedine, A., Albuquerque, A.L., and Da Silva Dias, P.L. 2006. Paleo-environmental change in Amazonian and African rainforest during the LGM. Palaeogeography, Palaeoclimatology, Palaeoecology, 239: 510-527.

Berthaud, J. 1986. Les ressources génétiques pour l'amélioration des caféiers africains diploïdes: évaluation de la richesse génétique des populations sylvestres et de ses mécanismes organisateurs, conséquences pour l'application. ORSTOM, Paris.

Charrier, A., and Berthaud, J. 1985. Botanical classification of coffee. *In* Coffee: Botany, Biochemistry and Production of Beans and Beverage. *Edited by* M.N. Clifford and M.N. Wilson. Croom Helm, London, pp. 13-47.

Combes, M.C., Andrzejewski, S., Anthony, F., Bertrand, B., Rovelli, P., Graziosi, G., and Lashermes, P. 2000. Characterization of microsatellite loci in *Coffea arabica* and related coffee species. Mol Ecol, 9(8): 1178-1180.

Cubry, P., De Bellis, F., Pot, D., Musoli, P., Legnaté, H., Leroy, T., and Dufour, M. 2005. Genetic diversity analyses and linkage disequilibrium evaluation in some natural and cultivated populations of *Coffea canephora*. *In* Proceedings of the 4th Plant genomics European meeting. Amsterdam, 20-23 september 2005.

Cubry, P., Musoli, P., Legnaté, H., Pot, D., De Bellis, F., Poncet, V., Anthony, F., Dufour, M., and Leroy, T. 2008 Diversity in coffee using SSR markers: structure of the Coffea genus and perspectives for breeding. Genome, 51 (1): 50-63.

Dufour, M., Hamon, P., Noirot, M., Risterucci, A.M., and Leroy, T. 2002. Potential use of SSR markers for *Coffea* spp. genetic mapping. *In* Proceedings of the 19th International Scientific colloquium on coffee, 2001-05-14/2001-05-18, Trieste, Italy. *Edited by* ASIC, Paris, France.

Dussert, S., Lashermes, P., Anthony, F., Montagnon, C., Trouslot, P., Combes, M-C., Berthaud, J., Noirot, M. and Hamon, S. 2003. Coffee (*Coffea canephora*) *In* Genetic Diversity of Cultivated Tropical Plants. *Edited by* P. Hamon, M. Seguin, X. Perrier and J. C. Glaszmann. Science Publishers. Inc. Enfield (NH), Plymouth. pp.239-258.

Evanno, G., Regnaut, S., and Goudet, J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Molecular Ecology, 14: 2611-2620.

Excoffier L., Laval, G., and Schneider, S. 2005. Arlequin ver. 3.0: An integrated software package for population genetics data analysis. Evolutionary Bioinformatics Online, 1: 47-50.

Gagiotti, O.L., Lange, O., Rassmann, K., and Gliddon, C. 1999. A comparison of two indirect methods for estimating average levels of gene flow using microsatellite data. Molecular Ecology, 8: 1513-1520.

Glaubitz, J.C. 2004. CONVERT: A user friendly program to reformat diploid genotypic data for commonly used population genetic software packages. Molecular Ecology Notes, 4: 309-310.

Goudet, J. 2001. FSTAT, a program to estimate and test gene diversities and fixation indices (version 2.9.3). Available from http://www.unil.ch/izea/softwares/fstat.html (updated from Goudet, 1995).

Goudet, J., Raymond, M., de Meeus, T., and Rousset, F. 1996. Testing differentiation in diploid populations. Genetics, 144: 1933-1940.

Hamilton, A.C. 1976. The significance of patterns of distribution shown by forest plants and animals in tropical Africa for the reconstruction of upper Pleistocene palaeoenvironments: a review. Palaeoecology of Africa, 9: 63-97.

Jolly, D., Taylor, D., Marchant, R., Hamilton, A., Bonnefille, R., Buchet, G., and Riollet, G. 1997. Vegetation dynamics in central Africa since 18,000 yr BP: pollen records from the interlacustrine highlands of Burundi, Rwanda and western Uganda. Journal of Biogeography, 24: 495-512.

Leakey, C.L.A. 1970. The Improvement of Robusta Coffee in East Africa. *In* Crop Improvement in East Africa. *Edited by* C.L. Leakey. Government Press, Kampala, pp. 250-277.

Leroy, T., Montagnon, C., Charrier, A., and Eskes, A.B. 1993. Reciprocal recurrent selection applied to *Coffea canephora* Pierre in Côte d'Ivoire. I Characterization and evaluation of breeding populations and value of intergroup hybrids. Euphytica, 67: 113-125.

Leroy, T., Marraccini, P., Dufour, M., Montagnon, C., Lashermes, P., Sabau, X., Ferreira, L.P., Jourdan, I., Pot, D., Andrade, A.C., Glaszmann, J.C., Vieira, L.G., and Piffanelli, P. 2005. Construction and characterization of a Coffea canephora BAC library to study the organization of sucrose biosynthesis genes. Theor Appl Genet, 111(6): 1032-1041.

Liu, K., and Muse, S.V., 2005. Powermarker: integrated analysis environment for genetic marker data. Bioinformatics, 21: 2128-2129.

Maitland, T.D. 1926. Coffea Robusta in Uganda. *In* Annual Conference of the Uganda planters' Association. *Edited by* Uganda Protectorate, Department of Agriculture. Government Press, Kampala, Uganda, pp. 3-11.

Maley, J. 1996. The African rain forest-main characteristics of changes. *In* Vegetation and climate from the upper Cretaceous to the Quaternary in Essays on the Ecology of the Guinea-Congo Rain Forest. *Edited by* I.J.Alexander, M.D. Swine and R. Watling. Royal Society of Edinburgh, Edinburgh, pp. 31-73.

Miller, J.C., and Tanksley, S.D. 1990. RFLP analysis of phylogenetic relationships and genetic variation in the genus Lycopersicon. Theoretical and Applied Genetics, 80: 437-448.

Montagnon, C., Leroy, T., and Yapo, A.B. 1992. Diversité génotypique et phénotypique de quelques groupes de caféiers (*Coffea canephora* Pierre) en collection. Conséquences sur leur utilisation en sélection. Café Cacao Thé, 36(3): 187-198.

Montagnon, C., Leroy, T., and Eskes, A.B. 1998. Varietal improvement of *Coffea canephora* II. Breeding programmes and their results. Plantation Recherche Développement, 5(2): 18-33.

Musoli, P. 2007. Recherche de sources de résistance à la trachéomycose du caféier *Coffea canephora* Pierre, due à *Fusarium xylarioides* Steyaert en Ouganda. PhD thesis, University of Montpellier, France.

Nei, M. 1973. Analysis of gene diversity in subdivided populations. Proceedings.of the National Academy of Sciences, 70: 3321-3323.

Nyakaana, S. 2007. Microgeographical genetic structure of forest Robusta coffee (*Coffea canephora*, Pierre), in Kibale National Park, Uganda. African Journal of Ecology, 45(1): 71-75.

Peakall, R., and Smouse, P. 2006. GENALEX6: genetic analysis in Excel. Population genetic software for teaching and research. Molecular Ecology Notes, 6: 288-295.

Perrier, X., Flori, A., and Bonnot, F. 2003. Data analysis methods. *In* Genetic diversity of cultivated tropical plants. *Edited by* P. Hamon, M. Seguin, X. Perrier, and J.C.Glaszmann. Inc. Enfield (NH), Plymouth, pp. 43–76.

Poncet, V., Hamon, P., Minier, J., Carasco, C., Hamon, S., and Noirot, M. 2004. SSR cross-amplification and variation within coffee trees (Coffea spp.). Genome, 47(6): 1071-1081.

Poncet, V., Dufour, M., Hamon, P., Hamon, S., De Kochko, A., and Leroy, T. 2007. Development of genomic microsatellite markers in Coffea canephora and their transferability to other coffee species. Genome, 50(12) 1156-1161.

Pritchard, J.K., Stephens, M., and Donnelly, P. 2000. Inference of population structure from multilocus genotype data. Genetics, 155: 945-959.

Risterucci, A.M., Grivet, L., N'Goran, J.A.K., Pierreti, I., Flament, M.H., and Lanaud, C. 2000. A high-density linkage map of *Theobroma cacao* L. Theoretical and Applied Genetics, 101(5/6): 948–955.

Rousset, F. 1997. Genetic differentiation and estimation of gene flow from F-Statistics under isolation by distance. Genetics, 147: 1219-1228.

Rovelli, P., Mettulio, R., Anthony, F., Anzueto, F., and Lashermes, P. 2000. Microsatellites in *Coffea arabica* L. *In* Coffee biotechnology and quality. *Edited by* T. Sera, C.R. Soccol, A. Pandey and S. Roussos. Kluwer Academic Publishers, The Netherlands. pp. 123-133.

Saitou, N., and Nei, M. 1987. The neighbour-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution, 4(4): 406-425.

Slatkin, M. 1995. A measure of population subdivision based on microsatellite allele frequencies. Genetics, 139: 457-462.

Thomas, A.S. 1935. Types of Robusta coffee and their selection in Uganda. The East African Agricultural Journa,l 1: 193-198.

Thomas, A.S. 1944. The wild coffees of Uganda. The Empire Journal of Experimental Agriculture, 12: 1-12.
Thomas, A.S. 1947. The cultivation and selection of Robusta coffee in Uganda. The Empire Journal of Experimental Agriculture, 15: 66-81.
Weir, B.S., and Cockerham, C.C. 1984. Estimating F-statistics for the analysis of populations' structure. Evolution, 38: 1358-1370.
White, F. 1979. The Guineo-Congolese region and its relationships to other phytochoria. Bulletin Du Jardin Botanique National de Belgique, 49: 11-55.

Table 1 - Microsatellite markers used for genotyping 232 *C. canephora* samples.

Marker code	LG[a]	Repeated sequence (SSR)	Allele size (bp)	Forward primer	Reverse primer	EMBL database number	Reference
351	3	$(GT)^{13}$	304	AAGGATGGCAAGTGGATTTCT	GCAGCTCTTGATTGTAGTTTCGT	AM23551	DUFOUR et al. 2002
355	1	$(TG)^{15}$	177	CTATGATGTCTTCCAACCTTCTAAC	GGTCCAATTCTGTTTCAATTTC	AM231552	DUFOUR et al. 2002
358	13	$(CA)^{11}$	248	CATGCACTATTATGTTTGTGTTTT	TCTCGTCATATTTACAGGTAGGTT	AM231554	DUFOUR et al. 2002
364	6	$(A)^{21}$	90	AGAAGAATGAAGACGAAACACA	TAACGCCTGCCATCG	AM231556	DUFOUR et al. 2002
368	8	$(TG)^{13}$	160	CACATCTCCATCCATAACCATTT	TCCTACCTACTTGCCTGTGCT	AM231558	DUFOUR et al. 2002
384		$(AC)^{10}$	255	ACGCTATGACAAGGCAATGA	TGCAGTAGTTTCACCCTTTATCC	AM231560	DUFOUR et al. 2002
394	6	$(TG)^{9}$	124	GCCGTCTCGTATCCCTCA	GAAGCCAGAAAGTCAGTCACATAG	AM231563	DUFOUR et al. 2002
429	2	$(A)^{13}$	175	CATTCGATGCCAACAACCT	GGGTCAACGCTTCTCCTG	AM231565	DUFOUR et al. 2002
442	8	$(CA)^{19}$	227	CGCAAATCTGAGTATCCCAAC	TGGATCAACACTGCCCTTC	AM231566	DUFOUR et al. 2002
445		$(AC)^{10}$	274	CCACAGCTTGAATGACCAGA	AATTGACCAAGTAATCACCGACT	AM231567	DUFOUR et al. 2002
456	8	$(AC)^{14}$	297	TGGTTGTTTTCTTCCATCAATC	TCCAGTTTCCCACGCTCT	AM231568	DUFOUR et al. 2002
461	2	$(AC)^{9}$	461	CGGCTGTGACTGATGTG	AATTGCTAAGGGTCGAGAA	AM231570	DUFOUR et al. 2002
463	10	$(AC)^{8}$	227	CATTCTTCCCACGATTCTATCTC	GTGACTTTCGGTTGAAATACTGG	AM231571	DUFOUR et al. 2002
471	1	$(CT)^{12}$	301	TTACCTCCCGGCCAGAC	CAGGAGACCAAGACCTTAGCA	AM231572	DUFOUR et al. 2002
501	7	$(TG)^{8}$	343	CACCACCATCTAATGCACCT	CTGCACCAGCTAATTCAAGC	AM231576	DUFOUR et al. 2002
753	6	$(CA)^{15}$	294	GGAGACGCAGGTGGTAGAAG	TCGAGAAGTCTTGGGGTGTT	AJ308753	ROVELLI *et al*. 2000
755	1	$(CA)^{20}$	184	CCCTCCCTCTTTCTCCTCTC	TCTGGGTTTTCTGTGTTCTCG	AJ250258	COMBES et al. 2000
774	3	$(CT)^{5}(CA)^{7}$	228	GCCACAAGTTTCGTGCTTTT	GGGTGTCGGTGTAGGTGTATG	AJ308774	ROVELLI *et al*. 2000
779	7	$(TG)^{17}$	116	TCCCCCATCTTTTTCTTTCC	GGGAGTGTTTTTGTGTTGCTT	AJ308779	ROVELLI *et al*. 2000
782	5	$(GT)^{15}$	114	AAAGGAAAATTGTTGGCTCTGA	TCCACATACATTTCCCAGCA	AJ308782	ROVELLI *et al*.2000
790	1	$(GT)^{21}$	134	TTTTCTGGGTTTTCTGTGTTCTC	TAACTCTCCATTCCCGCATT	AJ308790	ROVELLI *et al*. 2000
837	2	$(TG)^{16}(GA)^{1}$	102	CTCGCTTTCACGCTCTCTCT	CGGTATGTTCCTCGTTCCTC	AJ308837	ROVELLI *et al*. 2000
DL013	2	$(CA)^{6}(CT)8$	267	AGAGGGATGTCAGCATAA	ATTTGTGTTTGGTAGATGTG	AJ871892	LEROY et al. 2005
DL025	1	$(C)^{17}$	197	TTGTTGAGAGTGGAGGA	CCAAAGACAGTGCAGTAA	AJ871902	LEROY et al. 2005

[a] LG: Linkage group of the genetic map

Table 2 - Basic diversity statistics for *Coffea canephora* regions in Uganda

Region		No. of individuals	Total Number of Alleles	Mean Allele Number	Gene Diversity	Observed Heterozygosity	No of private alleles
Itwara	**Mean**[a]	56	193	8.062	0.586	0.396	37
	SD[a]			*0.929*	*0.041*	*0.044*	
	2.5% l.b.[a]			*6.333*	*0.502*	*0.310*	
	97.5% u.b[a]**.**			*9.958*	*0.665*	*0.482*	
Kibale	**Mean**	54	177	7.354	0.534	0.286	19
	SD			*0.834*	*0.051*	*0.043*	
	2.5% l.b.			*5.792*	*0.432*	*0.200*	
	97.5% u.b.			*9.000*	*0.637*	*0.373*	
Kalangala	**Mean**	35	206	8.626	0.628	0.406	34
	SD			*0.924*	*0.049*	*0.045*	
	2.5% l.b.			*6.875*	*0.530*	*0.315*	
	97.5% u.b.			*10.542*	*0.724*	*0.492*	
Erect	**Mean**	21	172	7.185	0.626	0.395	12
	SD			*0.757*	*0.048*	*0.045*	
	2.5% l.b.			*5.708*	*0.531*	*0.310*	
	97.5% u.b.			*8.625*	*0.716*	*0.488*	
Nganda	**Mean**	31	194	8.096	0.626	0.408	12
	SD			*0.772*	*0.048*	*0.049*	
	2.5% l.b.			*6.625*	*0.531*	*0.313*	
	97.5% u.b.			*9.667*	*0.713*	*0.501*	

Table 3 - Pairwise F_{ST} coefficients among *Coffea canephora* populations within Uganda

		Itwara					Kalangala					Kibale				Nganda
Region	**Site**	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	
Erect		0.17[a]ns	**0.20**	0.19ns	0.23ns	**0.20**	**0.11**	**0.09**	**0.14**	0.12ns	0.12ns	**0.29**	**0.17**	0.20ns	0.24ns	0.03ns
Itwara	1	0	0.01ns	0.01ns	0.05ns	0.01ns	**0.16**	**0.15**	**0.20**	0.17ns	0.15ns	**0.24**	**0.16**	0.10ns	0.18ns	0.13ns
	2		0	0.01ns	0.04ns	0.01ns	**0.17**	**0.16**	**0.20**	**0.18**	**0.15**	**0.20**	**0.17**	**0.10**	**0.22**	**0.13**
	3			0	0.03ns	0.00ns	**0.16**	**0.16**	**0.19**	0.18ns	0.15ns	**0.26**	**0.18**	0.10ns	0.19ns	0.14ns
	4				0	0.04ns	**0.24**	**0.23**	**0.28**	0.23ns	0.21ns	**0.31**	**0.30**	0.15ns	0.29ns	0.18ns
	5					0	**0.18**	**0.17**	**0.21**	**0.19**	**0.16**	**0.22**	**0.18**	**0.12**	**0.21**	**0.16**
Kalangala	1						0	0.00ns	0.06ns	0.01ns	0.00ns	**0.36**	**0.26**	**0.20**	**0.30**	**0.08**
	2							0	0.08ns	0.03ns	0.04ns	**0.33**	**0.22**	**0.21**	**0.29**	**0.09**
	3								0	0.11ns	0.03ns	**0.36**	**0.28**	**0.23**	**0.30**	**0.12**
	4									0	0.04ns	**0.33**	**0.25**	0.22ns	0.29ns	0.10ns
	5										0	**0.26**	**0.19**	**0.18**	0.25ns	0.09ns
Kibale	1											0	**0.35**	**0.18**	**0.32**	**0.19**
	2												0	**0.16**	**0.21**	**0.14**
	3													0	**0.11**	**0.14**
	4														0	0.21ns

[a] Significance of P value for 1,000 permutations is indicated: significant values (p-values 5% threshold for 1,000 permutations) are shown in bold; ns=not significant.

Table 4 - F_{ST} pairwise cœfficients between *Coffea canephora* regions in Africa including Uganda

	Nganda	**Kibale**	**Itwara**	**Kalangala**	**B**	**SG2**	**SG1**	**C**	**G**
Erect	0.04[a]ns	0.24*	0.23*	0.11*	0.38*	0.19ns	0.48*	0.37*	0.33ns
Nganda	0	0.16*	0.16*	0.07*	0.35*	0.22ns	0.52*	0.39*	0.34ns
Kibale		0	0.10*	0.21*	0.42*	0.33*	0.59*	0.49*	0.43*
Itwara			0	0.18*	0.37*	0.28*	0.52*	0.43*	0.39*
Kalangala				0	0.33*	0.17ns	0.46*	0.35*	0.31*
B					0	0.29*	0.63*	0.46*	0.42*
SG2						0	0.33*	0.21ns	0.24ns
SG1							0	0.37*	0.44*
C								0	0.31*

[a]Significance of P value for 1,000 permutations is indicated, ns=not significant; *=significant ($p<0.05$)

Table 5 – Isolation by distance pattern between all genetic groups and within Ugandan wild regions

	R value	P_{Mantel}[a]
Ugandan wild populations	0.411	0.015
All regions	0.497	0.132

[a] Probability for 9,999 permutations

Guinean genotypes, isolated from other groups, were not integrated in this study

Table 6 - Analysis of molecular variance (AMOVA)

AMOVA design	Source of Variance	Df[a]	sm[b]	Variance components	Percentage of explained variance
6 groups: Wild from Uganda, Guinean, Congolese B, C, SG1 & SG2	Among groups	5	298	2.23 (<0.01[c])	36.2
	Among individuals				
	Within groups	140	754	1.45(<0.01)	23.4
	Within individuals	146	365	2.50(<0.01)	40.4
7 groups: Cultivated and feral from Uganda, wild from Uganda, Guinean, B, C, SG1 and SG2.	Among groups	6	413	1.09(<0.01)	20.3
5 populations within Uganda: Erect, N'ganda, Kibale, Itwara and Kalangala					
	Among populations within groups	3	97	0.37(<0.01)	6.98
	Among individuals within populations	223	1131	1.16(<0.01)	21.5
	Within individuals	233	643	2.76(<0.01)	51.3
5 groups within Uganda corresponding to regions	Among groups	4	287	0.85(<0.01)	13.5
	Among individuals within groups	192	1289	1.25(<0.01)	19.9
	Within individuals	197	829	4.21(<0.01)	66.7

[a] Df = degree of freedom; [b] sm = sum of squares. [c] P value for variance components is indicated in brackets

Figure captions

Figure 1 - Geographical distribution of *C. canephora* Pierre diversity groups. The Guinean region covers Ivory Coast and Guinea. The Congolese regions (B, C, SG1, and SG2) are located in the Central African Republic, Cameroon and the Congo basin. ? is an unexplored region located in Angola. The map of Uganda shows the three surveyed regions: Kibale and Itwara forests, Kalangala islands.

Figure 2 - Hierarchical representation of sampling: country, regions and sites are identified. The number of individuals per region and site (N) is indicated in the boxes. The sampling area for Ugandan forests is indicated in brackets.

Figure 3 – Diversity tree of 196 individuals from Uganda (simple matching distance based on 24 microsatellite loci, weighted neighbour joining). I = Itwara forest; Ki = Kibale forest; N = Nganda, E = Erect; Ka = Kalangala islands.

Figure 4 - *F* statistics within Ugandan regions and sites. *FIS* values and significance for regions and sites are indicated in the corresponding boxes. F_{IT}, F_{ST} and F_{IS} are indicated for Ugandan regions including different sites, wild Ugandan genotypes and the country of Uganda. N indicates the sample size.

Figure 5 - WPGMA tree based on the Fst/(1-Fst) matrix (linearized Fst) of genetic distances between populations throughout Africa.

Figure 6 - Result of the population structure analysis within the country of Uganda. Four populations were identified: population A contains the Nganda (1) and Erect (2) regions. Population B is the Kibale forest (3), population C is the Itwara forest (4) and population D is Kalangala (5). Admixtures are presented by heterogeneous colours within populations.

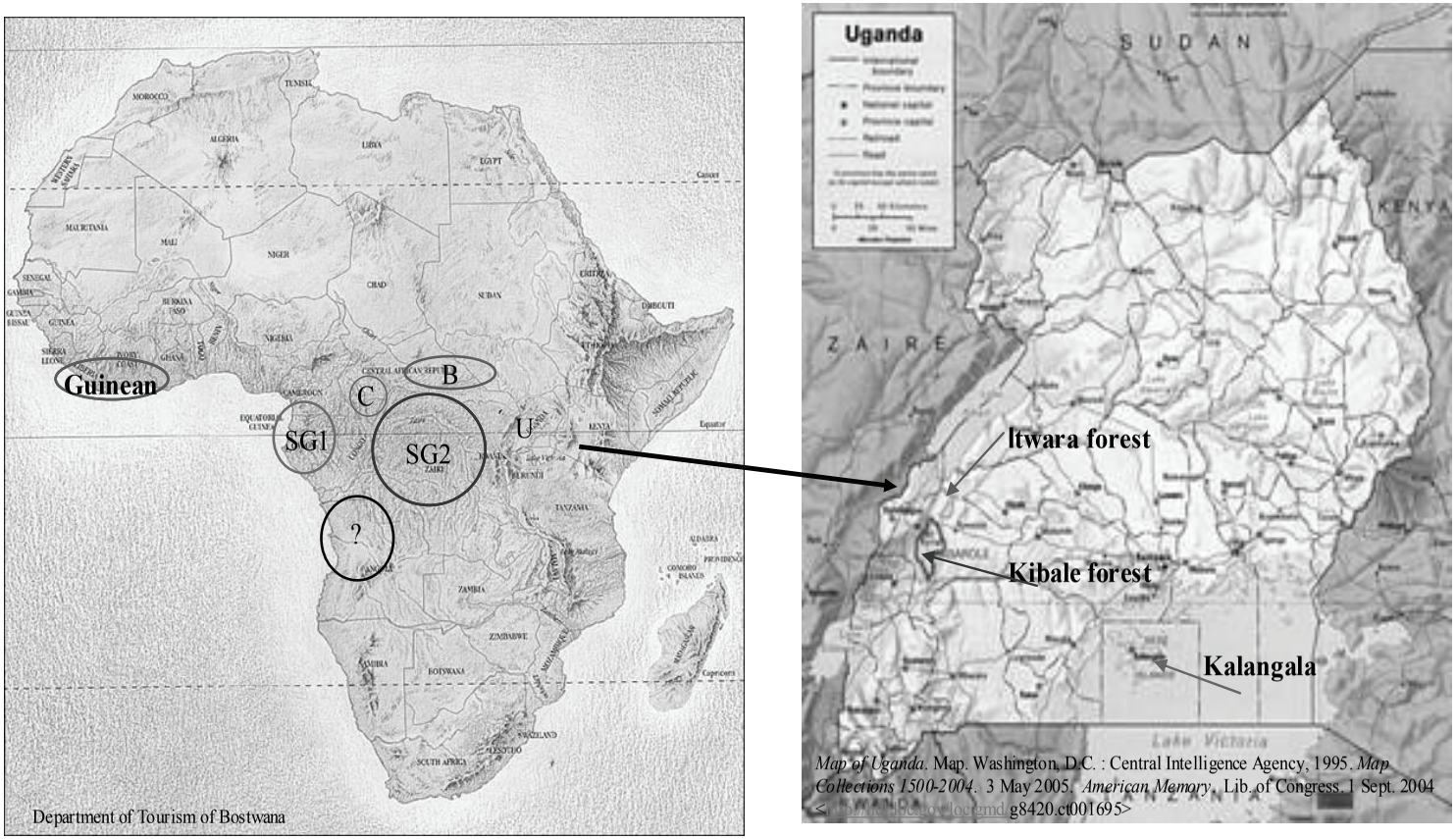

Guinean
B
C
U
SG1
SG2
?
Department of Tourism of Bostwana
Uganda
SUDAN
KENYA
ZAIRE
Itwara forest
Kibale forest
Kalangala
Lake Victoria
Map of Uganda. Map. Washington, D.C. : Central Intelligence Agency, 1995. Map Collections 1500-2004. 3 May 2005. American Memory. Lib. of Congress. 1 Sept. 2004 <

Figure 1

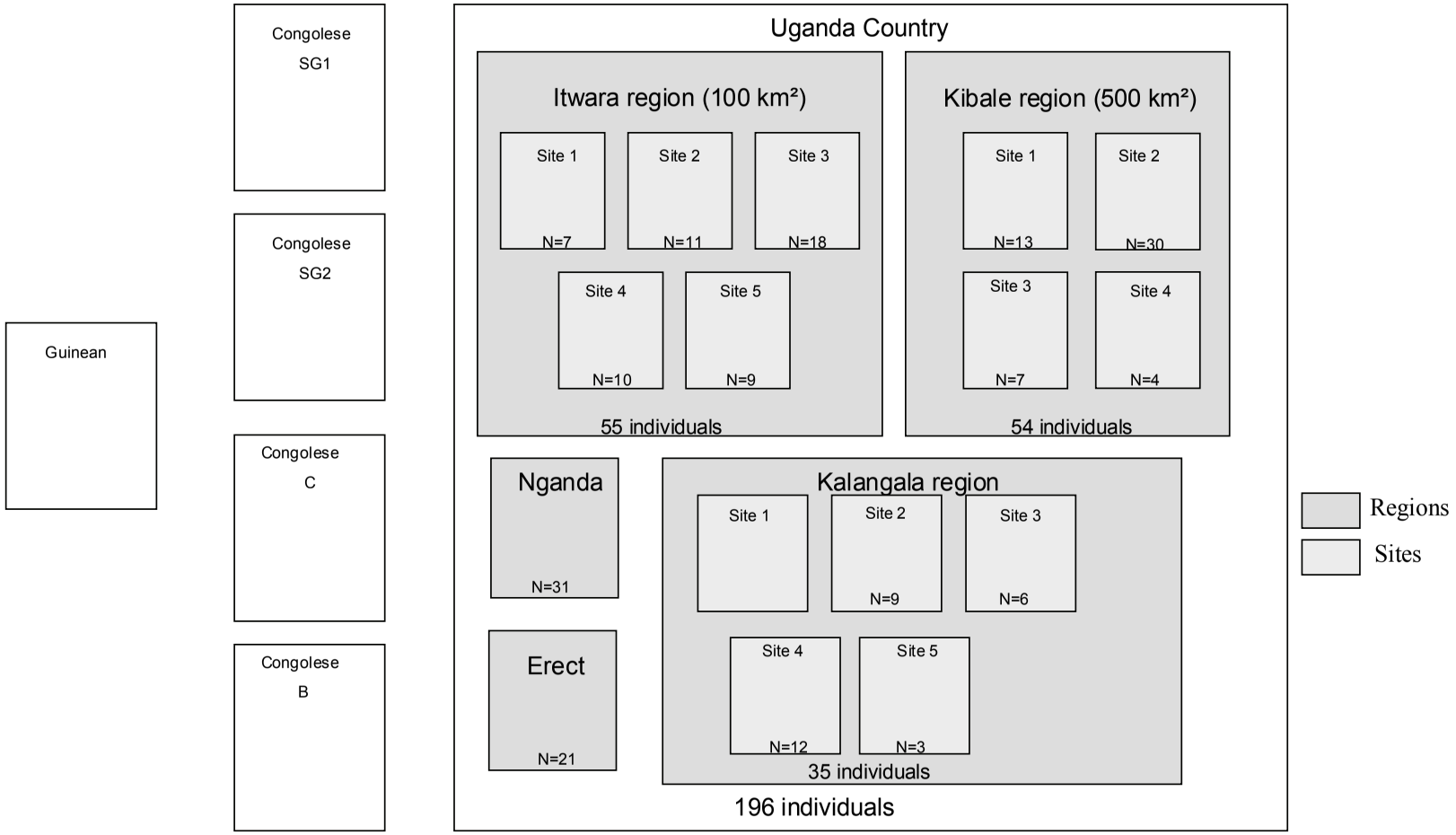

Congolese
SG1
Congolese
SG2
Guinean
Congolese
C
Congolese
B
Uganda Country
Itwara region (100 km²)
Site 1
N=7
Site 2
N=11
Site 3
N=18
Site 4
N=10
Site 5
N=9
55 individuals
Kibale region (500 km²)
Site 1
N=13
Site 2
N=30
Site 3
N=7
Site 4
N=4
54 individuals
Nganda
N=31
Erect
N=21
Kalangala region
Site 1
Site 2
N=9
Site 3
N=6
Site 4
N=12
Site 5
N=3
35 individuals
196 individuals
Regions
Sites

Figure 2

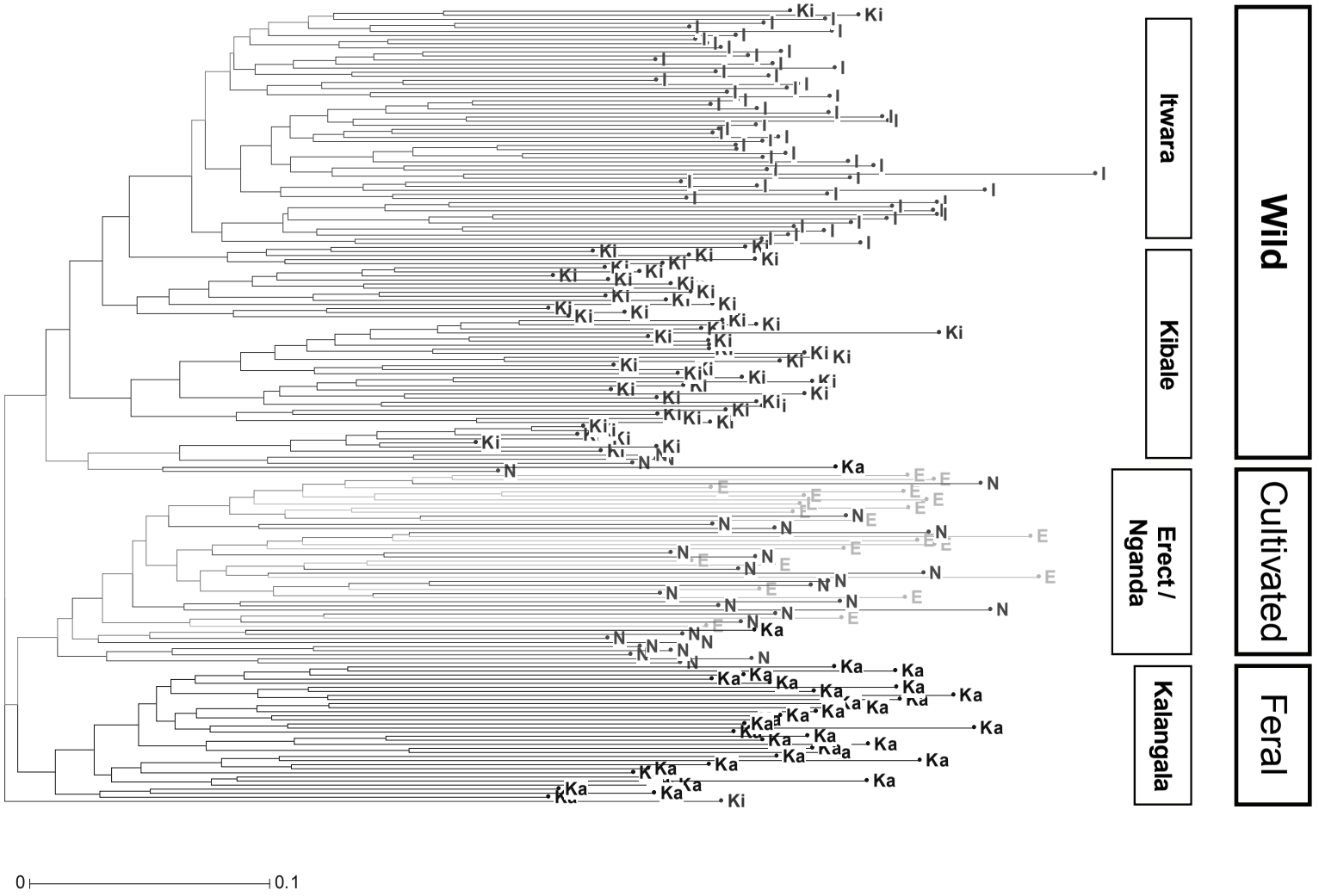

Wild
Cultivated
Feral
Itwara
Kibale
Erect / Nganda
Kalangala
0
0.1

Figure 3

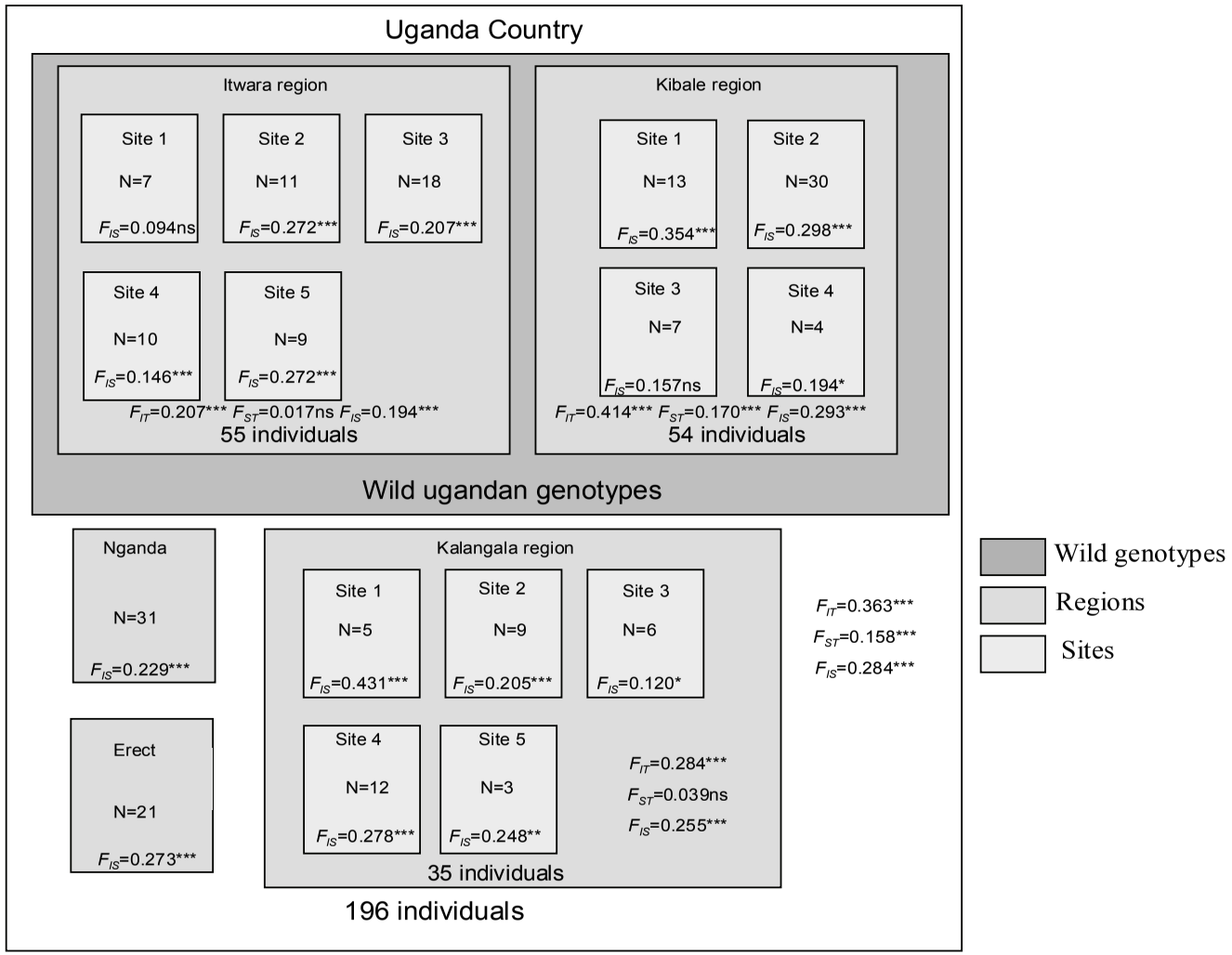

Uganda Country
Itwara region
Site 1
N=7
F_{IS}=0.094ns
Site 2
N=11
F_{IS}=0.272***
Site 3
N=18
F_{IS}=0.207***
Site 4
N=10
F_{IS}=0.146***
Site 5
N=9
F_{IS}=0.272***
F_{IT}=0.207*** F_{ST}=0.017ns F_{IS}=0.194***
55 individuals
Kibale region
Site 1
N=13
F_{IS}=0.354***
Site 2
N=30
F_{IS}=0.298***
Site 3
N=7
F_{IS}=0.157ns
Site 4
N=4
F_{IS}=0.194*
F_{IT}=0.414*** F_{ST}=0.170*** F_{IS}=0.293***
54 individuals
Wild ugandan genotypes
Nganda
N=31
F_{IS}=0.229***
Erect
N=21
F_{IS}=0.273***
Kalangala region
Site 1
N=5
F_{IS}=0.431***
Site 2
N=9
F_{IS}=0.205***
Site 3
N=6
F_{IS}=0.120*
Site 4
N=12
F_{IS}=0.278***
Site 5
N=3
F_{IS}=0.248**
F_{IT}=0.284***
F_{ST}=0.039ns
F_{IS}=0.255***
35 individuals
196 individuals
F_{IT}=0.363***
F_{ST}=0.158***
F_{IS}=0.284***
Wild genotypes
Regions
Sites

Figure 4

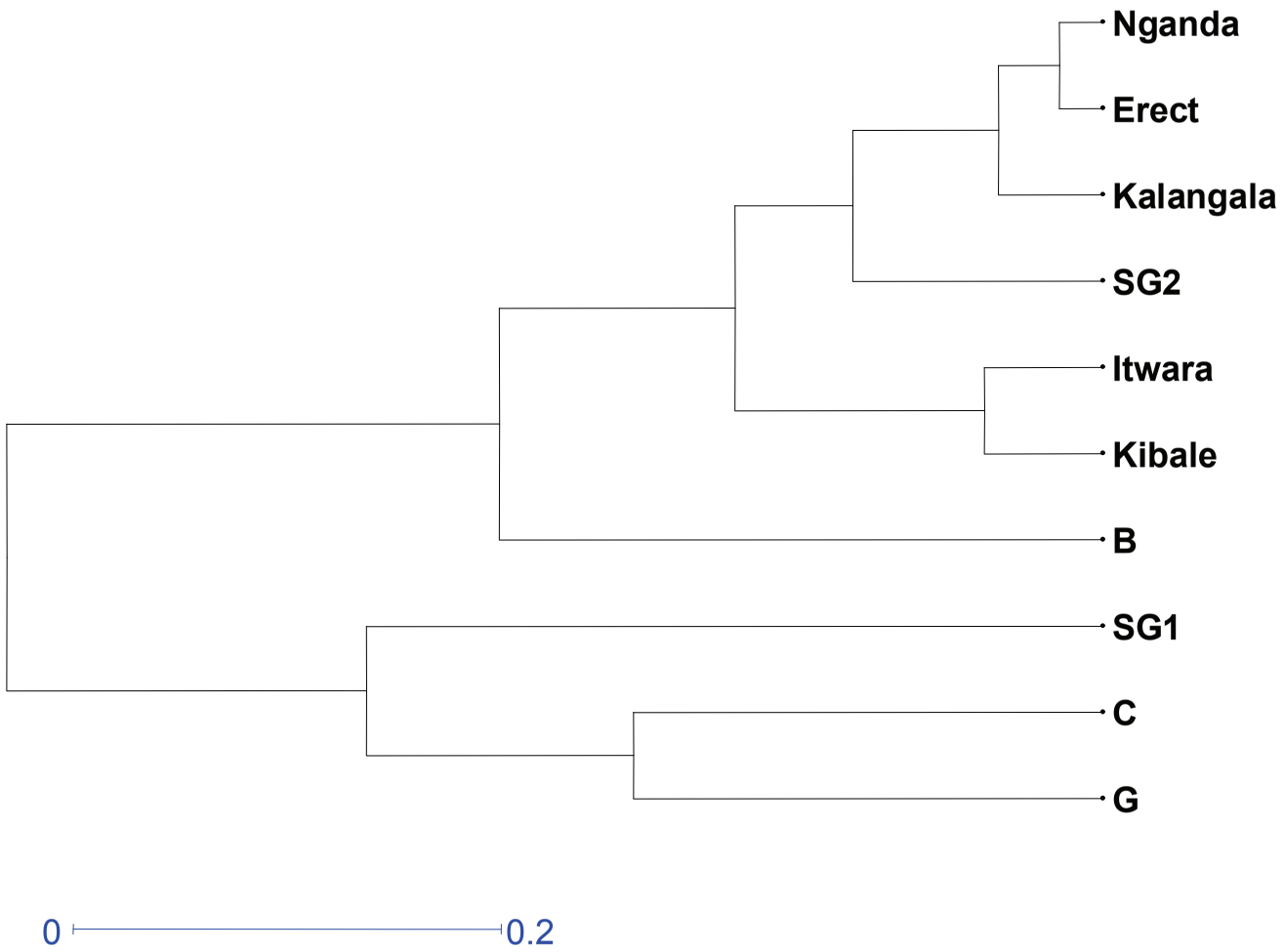


Figure 5

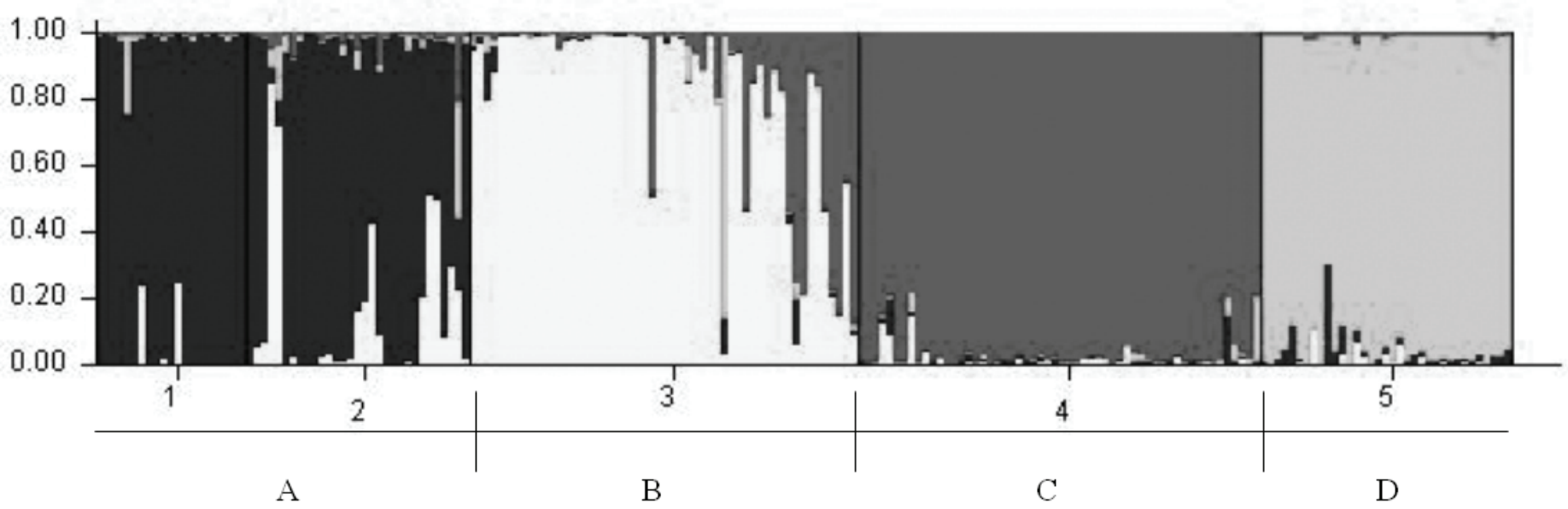
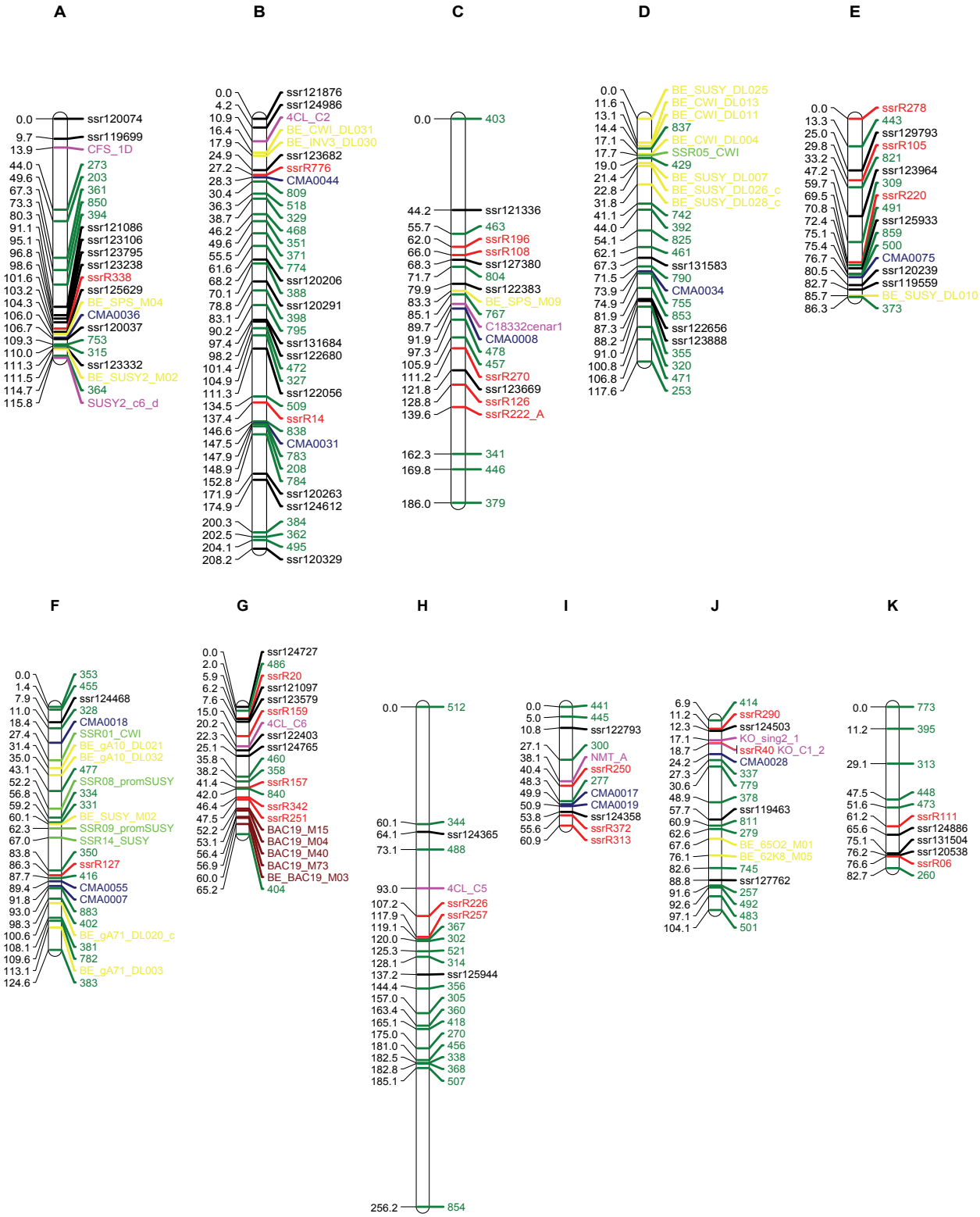


Figure 6

A.3.6 : Carte génétique de Coffea canephora développée au CIRAD au 31 octobre 2008



Carte génétique de *Coffea canephora* au 31 octobre 2008. Cette carte repose sur une descendance de 254 individus d'un croisement de type test-cross. Les marqueurs en verts foncé sont des microsatellites issus de banques génomiques, de même que les marqueurs en rouge et en bleu. Ces derniers (rouges et bleus) ont été fournis par d'autres équipes et pourront servir de ponts entre les diverses cartes génétiques. En noir on observe des marqueurs microsatellites issus de séquences EST fournis par Nestlé et qui pourront également servir de ponts. Les marqueurs en jaunes sont des microsatellites développés dans des séquences terminales de clone BAC ayant hybridés avec des gènes d'intérêt. En vert clair sont représentés des marqueurs microsatellites développés dans des séquences de gènes d'intérêt, enfin en violet sont représentés des polymorphismes d'insertion/délétion présent dans des séquences de gène candidats.

Chapitre 4

Liste des 356 génotypes utilisés dans cette étude

A.4.1 : Liste des génotypes utilisés dans cette étude

Ind	Groupe	pop
1635-1	SG1	Luki
1637-1	SG1	Luki
1640-1	SG1	Luki
1645-1	SG1	Luki
1647-1	SG1	Luki
1648-1	SG1	Luki
1649-1	SG1	Luki
br28	SG2	SG2
br29	SG2	SG2
br31	SG2	SG2
br32	SG2	SG2
c1016	C	C
c2001	C	C
c2002	C	C
c2003	SG2	SG2
c2004	SG2	SG2
c2007	Hybride	Hybride
c2008	C	C
c2011	SG2	SG2
c2012	SG2	SG2
c2013	SG2	SG2
c2014	SG2	SG2
c2015	C	C
c2016	SG2	SG2
c2017	SG2	SG2
c2018	SG2	SG2
c3001	SG1	Niaouli
c3002	SG1	Niaouli
c3003	SG1	Niaouli
c3004	SG1	Niaouli
c3005	SG1	Niaouli
c3006	SG1	Niaouli
c3007	SG1	Niaouli
c3008	SG1	Niaouli
c3009	SG1	Niaouli
c4001	C	Nana
c4002	C	Nana
c4003	C	Nana
c4004	C	Nana
c4005	C	Nana
c4006	C	Nana
c4007	C	Nana
c4008	C	Nana
c4009	C	Nana
c4010	C	Nana
c4011	C	Nana
c4012	C	Nana
c4013	C	Nana
c4014	C	Nana
c4015	C	Nana
c4016	C	Nana
c4017	C	Nana
c4018	C	Nana
c4019	C	Nana
c4020	C	Nana
c4021	C	Nana
c4022	C	Nana
c4023	C	Nana
c4024	C	Nana

c4025	C	Nana
c4026	C	Nana
c4027	C	Nana
c4028	C	Nana
c4029	C	Nana
c4030	C	Nana
c4031	C	Nana
c4032	C	Nana
c4033	C	Nana
c4034	C	Nana
c4035	C	Nana
c4036	C	Nana
c4037	C	Nana
c4038	C	Nana
c4039	C	Nana
c4040	C	Nana
c4041	C	Nana
c4042	C	Nana
c4043	C	Nana
c4044	C	Nana
c4045	C	Nana
c4046	C	Nana
c4047	C	Nana
c4048	C	Nana
c4049	C	Nana
c4050	C	Nana
c4051	C	Nana
c4052	C	Nana
c4053	C	Nana
c4054	C	Nana
c4055	C	Nana
c4056	C	Nana
c4057	C	Nana
c4058	C	Nana
c4059	C	Nana
c4060	C	Nana
c4061	C	Nana
c4062	C	Nana
c4063	C	Nana
c4064	C	Nana
c4065	C	Nana
c4066	C	Nana
c4067	C	Nana
c4068	C	Nana
c4069	C	Nana
c4070	C	Nana
c4071	C	Nana
c4072	C	Nana
c4073	C	Nana
c4074	C	Nana
c4075	C	Nana
c4076	C	Nana
c4077	C	Nana
c4078	C	Nana
c4079	C	Nana
c4080	C	Nana
c4081	C	Nana
c4082	C	Nana
c4083	C	Nana
c4084	C	Nana
c4085	C	Nana
c5001	SG2	Ineac7
c5002	SG2	Ineac7
c5003	SG2	Ineac7
c5004	SG2	Ineac7
c5005	SG2	Ineac7
c5006	SG2	Ineac7
c5007	SG2	Ineac7
g1001	Pelezi	Pelezi
g1002	Pelezi	Pelezi
g1003	Pelezi	Pelezi
g1004	Pelezi	Pelezi
g1005	Pelezi	Pelezi
g1006	Pelezi	Pelezi
g1007	Pelezi	Pelezi
g1008	Pelezi	Pelezi
g1009	Pelezi	Pelezi
g1010	Pelezi	Pelezi

g1011	Hybride	Hybride
g1013	Pelezi	Pelezi
g1014	Pelezi	Pelezi
g1015	Pelezi	Pelezi
g1016	Pelezi	Pelezi
g1017	Pelezi	Pelezi
g1018	Pelezi	Pelezi
g1019	Pelezi	Pelezi
g1020	Pelezi	Pelezi
g1021	Pelezi	Pelezi
g1022	Pelezi	Pelezi
g1023	Pelezi	Pelezi
g1024	Pelezi	Pelezi
g1025	Pelezi	Pelezi
g1026	Pelezi	Pelezi
g1027	Pelezi	Pelezi
g1028	Pelezi	Pelezi
g1029	Pelezi	Pelezi
g1030	Pelezi	Pelezi
g1031	Pelezi	Pelezi
g1032	Pelezi	Pelezi
g1033	Pelezi	Pelezi
g1034	Pelezi	Pelezi
g1035	Pelezi	Pelezi
g1036	SG2	SG2
g1037	Pelezi	Pelezi
g1038	Pelezi	Pelezi
g2001	G	pine
g2002	G	pine
g2003	G	pine
g2004	G	Hybride
g2005	G	pine
g2006	G	pine
g2009	G	pine
g2010	G	pine
g2012	G	pine
g2013	G	pine
g2014	G	pine
g2015	G	pine
g2016	G	Hybride
g2017	G	pine
g2018	G	pine
g2019	G	pine
g2020	G	pine
g2021	G	pine
g2023	G	pine
g2024	G	pine
g2025	G	pine
g2026	G	pine
g2027	G	pine
g2028	G	pine
g2029	G	pine
g2030	G	pine
g2031	G	pine
g2032	G	pine
g2033	G	pine
g2034	G	pine
g2035	G	pine
g3001	G	guincultiv
g3002	G	guincultiv
g3004	G	guincultiv
g3005	G	guincultiv
g3007	G	guincultiv
g3008	G	guincultiv
g3009	G	guincultiv
g3010	G	guincultiv
g3011	G	guincultiv
g3012	G	guincultiv
g3013	G	guincultiv
g3015	SG2	SG2
g3017	G	guincultiv
g3018	G	guincultiv
g3019	G	guincultiv
g3020	G	guincultiv
g3021	G	guincultiv
g4001	G	ira1
g4002	G	ira1
g4003	G	ira1

g4005	G	ira1	g6016	G	fourougbankoro
g4006	G	ira1	g6017	G	fourougbankoro
g4007	G	ira1	g6018	G	fourougbankoro
g4008	G	ira1	g6019	G	fourougbankoro
g4009	G	ira1	g6020	G	fourougbankoro
g5001	G	ira2	g6021	G	fourougbankoro
g5002	G	ira2	g7001	G	mouniandougou
g5003	G	ira2	g7002	G	mouniandougou
g5004	G	ira2	g7003	G	mouniandougou
g5005	G	ira2	g7004	G	mouniandougou
g5006	G	ira2	g7005	G	mouniandougou
g5007	G	ira2	g7006	G	mouniandougou
g5008	G	ira2	g7007	G	mouniandougou
g5009	G	ira2	g7008	G	mouniandougou
g5010	G	ira2	g7009	G	mouniandougou
g5011	G	ira2	g7010	G	mouniandougou
g5012	G	ira2	g7011	G	mouniandougou
g5013	G	ira2	g7012	G	mouniandougou
g5014	G	ira2	g7013	G	mouniandougou
g5015	G	ira2	g7014	G	mouniandougou
g5016	Hybride	Hybride	g7015	G	mouniandougou
g5017	G	ira2	g7016	G	mouniandougou
g5018	G	ira2	g7017	G	mouniandougou
g5019	G	ira2	g7018	G	mouniandougou
g6001	G	fourougbankoro	g7019	G	mouniandougou
g6002	G	fourougbankoro	g7020	G	mouniandougou
g6003	G	fourougbankoro	g7021	G	mouniandougou
g6004	G	fourougbankoro	g7022	G	mouniandougou
g6005	G	fourougbankoro	g7023	G	mouniandougou
g6006	G	fourougbankoro	g7024	G	mouniandougou
g6007	G	fourougbankoro	g7025	G	mouniandougou
g6008	G	fourougbankoro	g7026	G	mouniandougou
g6009	G	fourougbankoro	g7027	G	mouniandougou
g6010	G	fourougbankoro	g7028	G	mouniandougou
g6011	G	fourougbankoro	g7029	G	mouniandougou
g6012	G	fourougbankoro	g7030	G	mouniandougou
g6013	G	fourougbankoro	g7031	G	mouniandougou
g6014	G	fourougbankoro	g7032	G	mouniandougou
g6015	G	fourougbankoro	g8001	G	sabregue

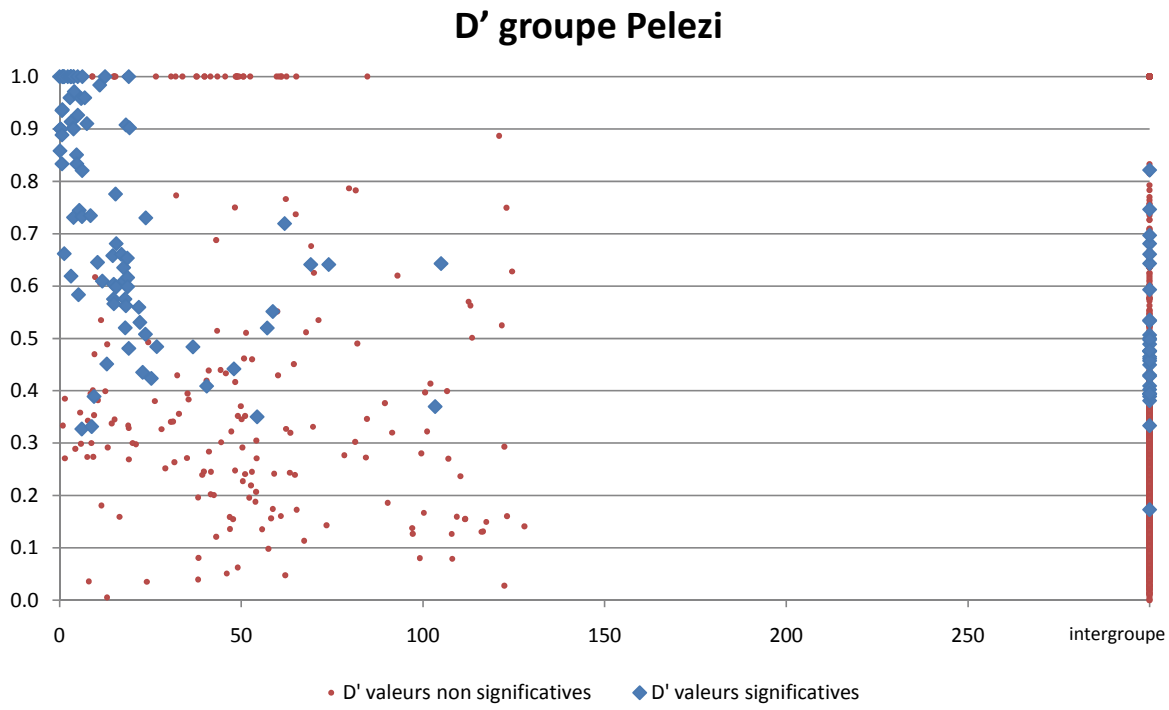
g8002	C	C
in002	SG2	SG2
in003	C	C
in004	G	inc
sp43	SG2	SG2
sp44	SG2	SG2
sp45	SG2	SG2
sp49	SG2	SG2
ue001	SG2	erect
ue002	SG2	erect
ue003	SG2	erect
ue005	SG2	erect
ue006	SG2	erect
ue007	SG2	erect
ue008	SG2	erect
ue009	SG2	erect
ue010	SG2	erect
ue011	SG2	erect
ue012	SG2	erect
ue013	SG2	erect
ue014	SG2	erect
ue015	SG2	erect
ue016	SG2	erect
ue017	SG2	erect
ue019	SG2	erect
ue020	SG2	erect
ue021	SG2	erect
ue025	SG2	erect
ue026	SG2	erect
ue027	SG2	nganda
ue029	SG2	nganda
ue030	SG2	nganda
un001	SG2	nganda
un002	SG2	nganda
un003	SG2	nganda
un004	SG2	nganda
un005	SG2	nganda
un006	SG2	nganda
un008	SG2	nganda
un009	SG2	nganda
un010	SG2	nganda
un011	SG2	nganda
un013	SG2	nganda
un014	SG2	nganda
un015	SG2	nganda
un017	SG2	nganda
un018	SG2	nganda
un019	SG2	nganda
un020	SG2	nganda
un022	SG2	nganda
un023	SG2	nganda
un024	SG2	nganda
un025	SG2	nganda
un027	SG2	nganda
un028	SG2	nganda
un029	SG2	nganda
un030	SG2	nganda
un031	SG2	nganda
un032	SG2	nganda
un033	SG2	nganda
un034	SG2	nganda
un035	SG2	nganda
un037	SG2	nganda

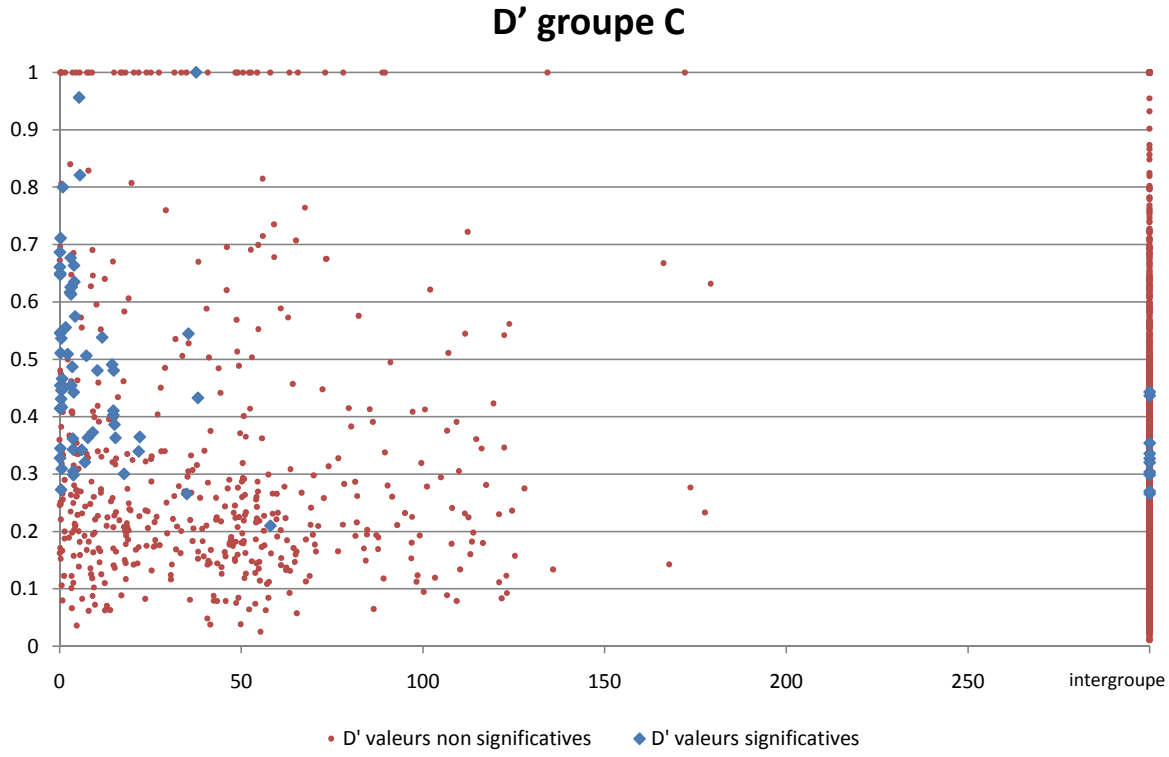
Décroissance de D' selon la distance génétique

A.4.2 : décroissance de D' en fonction de la distance génétique pour les groupes Pélézi et C.

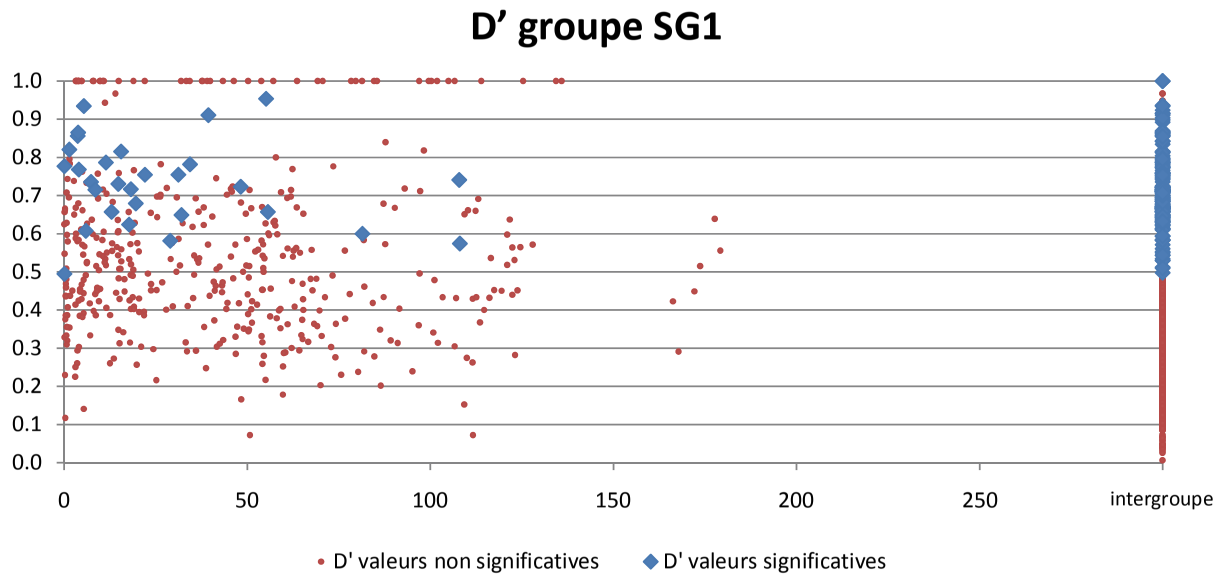
A.4.3 : décroissance de D' en fonction de la distance génétique pour les groupes SG1 et SG2.

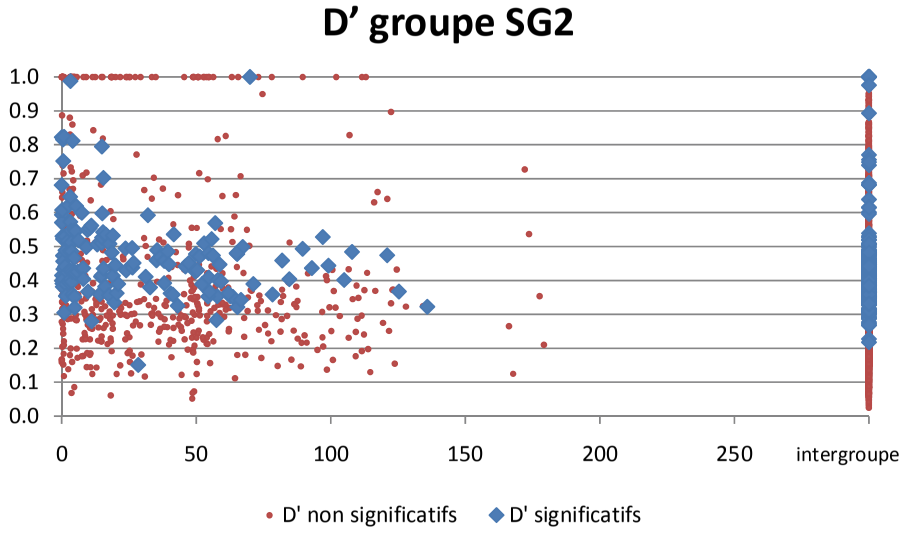
A.4.4 : décroissance de D' en fonction de la distance génétique pour le groupe G et les trois sous-groupes de G identifiés.

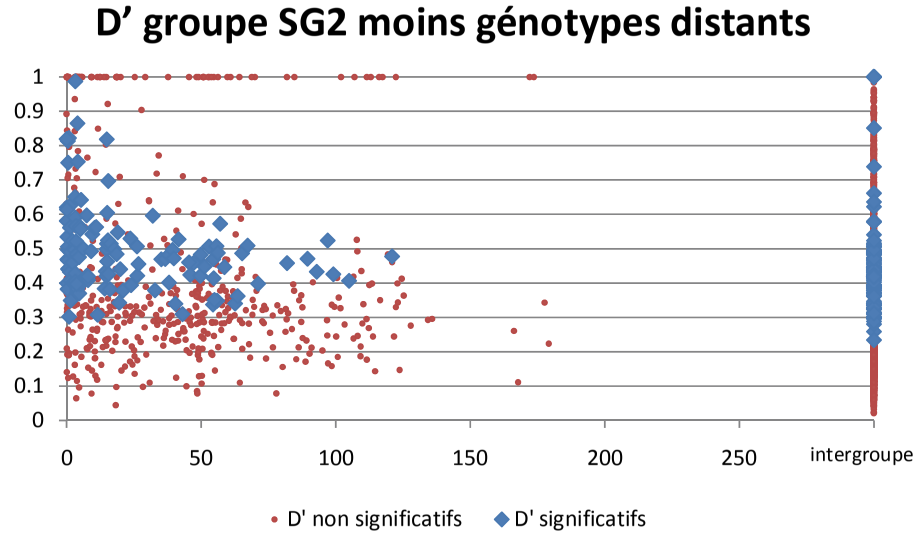




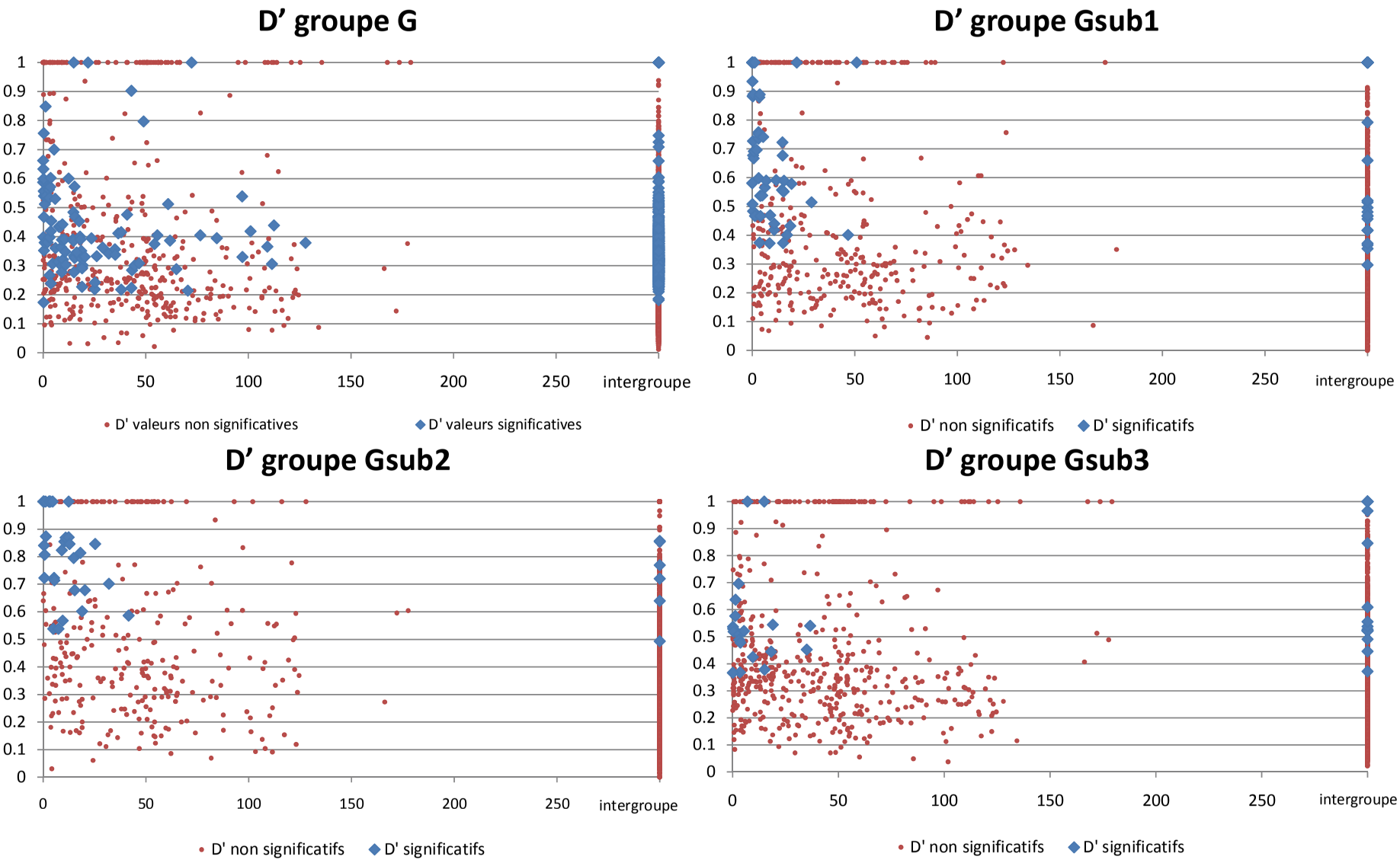
D’ en fonction de la distance génétique pour les groupes Pelezi (haut) et C (bas).







D’ en fonction de la distance génétique pour les groupes SG1 et SG2.



D’ en fonction de la distance génétique pour le groupe G et les trois sous-groupe identifiés.

Chapitre 5

Indices de diversité des polymorphismes du clone BAC 111O18

A.5.1 : Tableau descriptif des polymorphismes du clone BAC 111O18

	48 génotypes				G				C			
	Genotype		**Gene**		**Genotype**		**Gene**		**Genotype**		**Gene**	
Marker	**No**	**Allele No**	**Diversity**	**Obs Het**	**No**	**Allele No**	**Diversity**	**Obs Het**	**No**	**Allele No**	**Diversity**	**Obs Het**
SNP_B19_001	3	2	0.187	0.167	3	2	0.320	0.320	NA	NA	NA	NA
SNP_B19_002	3	2	0.061	0.021	NA	NA	NA	NA	3	2	0.122	0.043
SNP_B19_003	3	2	0.364	0.188	3	2	0.497	0.360	NA	NA	NA	NA
SNP_B19_004	2	2	0.021	0.021	2	2	0.039	0.040	NA	NA	NA	NA
SNP_B19_005	3	2	0.478	0.083	NA	NA	NA	NA	3	2	0.287	0.174
SNP_B19_006	3	2	0.234	0.104	3	2	0.385	0.200	NA	NA	NA	NA
SNP_B19_007	3	2	0.495	0.521	3	2	0.493	0.480	3	2	0.496	0.565
SNP_B19_008	3	2	0.437	0.438	3	2	0.343	0.360	3	2	0.491	0.522
SNP_B19_009	3	2	0.234	0.188	NA	NA	NA	NA	3	2	0.405	0.391
B19_M03	22	9	0.826	0.604	15	8	0.794	0.680	9	4	0.632	0.522
B19_M04	10	5	0.630	0.500	4	3	0.390	0.400	8	4	0.700	0.609
B19_M15	6	4	0.338	0.313	5	4	0.282	0.240	3	2	0.364	0.391
B19_M34	2	2	0.080	0.083	NA	NA	NA	NA	2	2	0.159	0.174
SNP_B19_010	3	2	0.413	0.250	NA	NA	NA	NA	3	2	0.476	0.522
SNP_B19_011	2	2	0.041	0.042	NA	NA	NA	NA	2	2	0.083	0.087
SNP_B19_012	2	2	0.080	0.000	2	2	0.147	0.000	NA	NA	NA	NA
SNP_B19_013	2	2	0.170	0.188	2	2	0.295	0.360	NA	NA	NA	NA
SNP_B19_015	3	2	0.492	0.417	2	2	0.295	0.360	3	2	0.405	0.478
SNP_B19_014	2	2	0.021	0.021	NA	NA	NA	NA	2	2	0.043	0.043
B19_M35	8	4	0.627	0.479	3	2	0.385	0.440	8	4	0.696	0.522
B19_M45	27	12	0.844	0.771	16	9	0.802	0.720	19	11	0.853	0.826
B19_M50	4	4	0.539	0.083	3	4	0.186	0.160	NA	NA	NA	NA
B19_M57	13	6	0.760	0.563	7	4	0.566	0.480	10	5	0.728	0.652
B19_M62	2	2	0.153	0.167	NA	NA	NA	NA	2	2	0.287	0.348
B19_M63	4	3	0.504	0.083	2	2	0.077	0.000	4	3	0.232	0.174
SNP_B19_016	3	2	0.234	0.229	NA	NA	NA	NA	3	2	0.405	0.478
SNP_B19_017	2	2	0.041	0.000	2	2	0.077	0.000	NA	NA	NA	NA
SNP_B19_018	2	2	0.021	0.021	2	2	0.039	0.040	NA	NA	NA	NA
SNP_B19_019	2	2	0.021	0.021	2	2	0.039	0.040	NA	NA	NA	NA
SNP_B19_020	2	2	0.021	0.021	2	2	0.039	0.040	NA	NA	NA	NA
SNP_B19_021	3	2	0.135	0.104	NA	NA	NA	NA	3	2	0.258	0.217
SNP_B19_022	3	2	0.500	0.021	2	2	0.077	0.000	2	2	0.043	0.043
SNP_B19_023	3	2	0.234	0.229	NA	NA	NA	NA	3	2	0.405	0.478
SNP_B19_024	3	2	0.353	0.292	NA	NA	NA	NA	3	2	0.499	0.609

SNP_B19_025	2	2	0.041	0.042	2	2	0.077	0.080	NA	NA	NA	NA
SNP_B19_026	3	2	0.135	0.021	3	2	0.241	0.040	NA	NA	NA	NA
SNP_B19_027	3	2	0.219	0.167	3	2	0.365	0.320	NA	NA	NA	NA
SNP_B19_028	3	2	0.500	0.021	2	2	0.077	0.000	2	2	0.043	0.043
SNP_B19_029	3	2	0.500	0.021	2	2	0.077	0.000	2	2	0.043	0.043
SNP_B19_030	3	2	0.153	0.083	3	2	0.269	0.160	NA	NA	NA	NA
SNP_B19_031	2	2	0.021	0.021	2	2	0.039	0.040	NA	NA	NA	NA
SNP_B19_032	3	2	0.500	0.021	2	2	0.147	0.000	2	2	0.043	0.043
SNP_B19_033	2	2	0.061	0.063	2	2	0.113	0.120	NA	NA	NA	NA
SNP_B19_034	2	2	0.041	0.042	2	2	0.077	0.080	NA	NA	NA	NA
SNP_B19_035	3	2	0.495	0.063	NA	NA	NA	NA	2	2	0.122	0.130
SNP_B19_036	3	2	0.498	0.021	NA	NA	NA	NA	2	2	0.043	0.043
SNP_B19_037	3	2	0.498	0.021	NA	NA	NA	NA	2	2	0.043	0.043
SNP_B19_038	3	2	0.498	0.021	NA	NA	NA	NA	2	2	0.043	0.043
SNP_B19_039	3	2	0.498	0.021	NA	NA	NA	NA	2	2	0.043	0.043
SNP_B19_040	3	2	0.249	0.208	3	2	0.403	0.400	NA	NA	NA	NA
B19_M66	15	7	0.765	0.375	8	4	0.502	0.320	7	4	0.595	0.435
B19_M73	14	9	0.781	0.292	7	6	0.575	0.400	7	6	0.604	0.174
B19_M75	2	2	0.499	0.000	NA	NA	NA	NA	NA	NA	NA	NA
B19_M79	4	3	0.458	0.146	4	3	0.215	0.160	3	2	0.485	0.130
BE_BAC19_M03	9	5	0.588	0.563	7	4	0.640	0.560	4	3	0.487	0.565
Mean	5	3	0.338	0.172	3.783	2.783	0.281	0.227	4.028	2.74	0.333	0.303

Obs Het : hétérozygotie observée, NA : marqueur non polymorphe

Summary:

The search for associations between molecular markers and variation of traits with agronomic importance is a main goal of Marker-Assisted Selection development, especially for perennials species. Knowledge of genetic diversity and structure is a prerequisite for such studies. In the same way, knowledge of structure and extant of Linkage Disequilibrium in populations is an important task.

We strengthen and fine-tuned previously described genetic structure and diversity of *Coffea canephora* Pierre in order to assess the capabilities of association mapping based approaches for this species. We laid the basis for rapid characterization of genetic resources with the help of microsatellite markers and clarify the genetic origin of a previously non-studied population. Future development of core-collections will benefit from these results.

Structure and extent of Linkage Disequilibrium was determined at different scales and for different populations or genetic groups, with the help of 108 genome wide markers, giving the opportunities to identify suitable populations for the 2 assocation mapping categories, genome-wide scan and candidate region.

A first try of association mapping confirms the important capabilities of such approaches for our species. Implications of genetic diversity and structure as well as Linkage Disequilibrium structure for future breeding purposes are discussed.

Keywords: *Coffea canephora* Pierre, genetic diversity, linkage disequilibrium, genetic structure, population genetics, association studies.

Résumé :

La recherche d'associations entre marqueurs moléculaires et les variations de caractères d'intérêt agronomique est un enjeu majeur du développement de la sélection assistée par marqueurs, notamment pour des espèces pérennes. La connaissance de la diversité génétique et de la structure de celle-ci est un préalable indispensable à la mise en place de telles études. Au même titre, la connaissance de l'étendue et de l'intensité du déséquilibre de liaison au sein des populations considérées est également nécessaire.

Nous avons donc entrepris de confirmer et de préciser la diversité et la structure génétique de *Coffea canephora* Pierre afin d'évaluer les potentialités d'études d'association sur cette espèce. Nous avons posé les bases d'une caractérisation rapide des collections de ressources génétiques de cette espèce à l'aide de marqueurs microsatellites et précisé l'origine génétique d'une population jusqu'à présent non étudiée. Cette base pourra également servir au développement de futures cores-collections.

L'évaluation de la structure et de l'étendue du déséquilibre de liaison à différentes échelles et au niveau de différentes populations ou groupes de diversité à l'aide de 108 marqueurs répartis sur l'ensemble du génome nous a permis de proposer des populations utilisables pour les 2 types d'approches de génétique d'associations, scan génome entier ou région candidate.

Un premier essai d'étude d'association a montré les potentialités importantes de ces approches pour notre espèce. Les implications de l'importance de la diversité génétique et de sa structure ainsi que de la structure du déséquilibre de liaison au sein de nos populations pour l'amélioration sont discutées.

Discipline : Evolution, ressources génétiques, génétique des populations

Mots-clefs : *Coffea canephora* Pierre, diversité génétique, déséquilibre de liaison, structure génétique, génétique des populations, études d'association.

Adresse du laboratoire où se sont déroulés les travaux de thèse :

CIRAD département BIOS, UMR DAP – TA A96-03

Avenue Agropolis 34093 Montpellier